Universidad
Carlos III de Madrid

Departamento de Teoría de la Señal y Comunicaciones

PROYECTO FIN DE CARRERA

RECONOCIMIENTO AUTOMÁTICO DE HABLA CON ADAPTACIÓN AL GÉNERO Y AL LOCUTOR

Autor: Ana Belén Caballero Pedrera

Tutor: Ascensión Gallardo Antolín

Director: Jesús de Vicente Peña

Leganés, diciembre de 2010

Título: Reconocimiento automático de habla con adaptación al género y al locutor
Autor: Ana Belén Caballero Pedrera
Director: Jesús de Vicente Peña
Tutor: Ascensión Gallardo Antolín

EL TRIBUNAL

Presidente: Fernando Díaz de María

Vocal: Mª Jesús Poza Lara

Secretario: Carmen Peláez Moreno

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día 16 de diciembre de 2010 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

Agradecimientos

Con la realización de este proyecto fin de carrera finaliza una etapa importante de mi vida. Finalizar mis estudios de Ingeniería de Telecomunicación no sólo me afectará profesionalmente, sino también, y lo que es más importante, personalmente, ya que significa la consecución de un reto que me propuse hace ya bastante tiempo.

En el camino recorrido hasta llegar a este momento, han influido muchas personas, y todas ellas merecen un agradecimiento sincero por mi parte.

En primer lugar, me gustaría agradecer a Fernando Díaz que me brindara la oportunidad de disfrutar de una beca en el Departamento de Teoría de la Señal y Comunicaciones, no sólo porque significó el inicio de mi proyecto fin de carrera, sino porque también me permitió descubrir el importante trabajo de investigación realizado en la Universidad Carlos III de Madrid.

Por supuesto, agradecer a Jesús de Vicente el gran esfuerzo invertido en la dirección de mi proyecto, por su ayuda técnica y por mantener ese soporte aún estando fuera de Madrid. Y gracias también a Ascensión Gallardo, al tomar el relevo de Jesús en estos últimos meses.

También merece mi sincero agradecimiento el Grupo de Procesado Multimedia del Departamento de Teoría de la Señal y Comunicaciones, formado por profesionales completamente involucrados con sus compañeros, y con quienes no lo son ya, con los que siempre he podido contar.

Por supuesto, acordarme de Salva, un buen compañero de beca, de estudios y amigo, al que siempre he podido acudir cuando lo he necesitado.

Por último, agradecer a todas las personas que forman parte de mi entorno más cercano y que me han acompañado durante todo este tiempo, apoyándome, ayudándome, animándome y, por supuesto, aguantándome en los momento más difíciles. Gracias por llegar hasta aquí conmigo, sin vosotros no hubiera sido posible.

Muchas gracias a todos.

Resumen

El presente proyecto fin de carrera tiene como objetivo mejorar las prestaciones de un reconocedor automático de habla continua en castellano, adaptando sus características al género y al locutor.

Para ello, como sistema base se ha desarrollado un reconocedor automático de habla continua para la base de datos MICROAES, utilizando el conjunto de herramientas de HTK con el que se realiza un modelado basado en HMMs.

Para observar el efecto que provoca introducir información sobre el género en el reconocimiento, se ha utilizado una adaptación MAP al género sobre los modelos acústicos previamente entrenados. Además, se ha implementado un clasificador de género con el fin de realizar automáticamente la elección del género correspondiente a cada locución. Finalmente, la adaptación MAP se ha comparado con el entrenamiento por separado de los modelos dependientes del género.

Para analizar la influencia producida por la incorporación de la información del locutor, se ha planteado el uso de dos técnicas de adaptación diferentes, MAP y MLLR.

Por último, se ha evaluado el uso conjunto de la adaptación al género y al locutor, utilizando una combinación de las técnicas de adaptación mencionadas anteriormente.

Palabras clave: reconocimiento automático de habla, HMMs, adaptación, MAP, MLLR, clasificador de género.

Abstract

This Final Degree Project aims to improve the recognition rate of an automatic speech recognition system in Spanish, by adapting its characteristics to gender and speaker variability.

For that, as a baseline system, an automatic continuous speech recognizer has been developed for the MICROAES database, using the HTK toolkit to perform a modeling based on HMMs.

To observe the effect caused by introducing gender information in the recognition system, a gender MAP adaptation of the acoustic models previously trained has been used. Furthermore, a gender classifier to automatically perform the choice of gender for each utterance has been implemented. Finally, the MAP adaptation technique has been compared to the separate training of gender dependent models.

To analyze the influence of the incorporation of the speaker information, two different adaptation techniques, MAP and MLLR, have been considered.

Finally, the joint use of gender and speaker adaptation has been evaluated using a combination of the adaptation techniques previously mentioned.

Keywords: automatic speech recognition, HMMs, adaptation, MAP, MLLR, gender classifier.

Índice general

1. INTRODUCCIÓN Y OBJETIVOS 1

1.1 Introducción 1

1.2 Objetivos 2

1.3 Esquema 3

1.4 Tecnologías del habla 4

1.4.1 Clasificación 4

1.4.2 Evolución histórica 5

1.4.3 Líneas futuras de desarrollo 7

2. RECONOCIMIENTO AUTOMÁTICO DE HABLA 9

2.1 Resumen 9

2.2 Introducción 10

2.3 Producción y recepción de voz 10

2.3.1 Generación de la señal de voz 11

2.3.2 Recepción de la señal de voz 14

2.3.3 Enfoque computacional 16

2.4 Reconocimiento automático de habla 17

2.4.1 Parametrización 20

2.4.1.1 Modelado fuente-filtro 20

2.4.1.2 Enventanado de la señal de voz 21

2.4.1.3 Procesamiento previo 22

2.4.1.4 Cepstrum 23

2.4.1.5 Análisis de banco de filtros....24
2.4.1.6 Análisis lineal predictivo....26
2.4.1.7 Técnicas a posteriori....27
2.4.2 Modelado acústico....27
2.4.2.1 Técnicas de reconocimiento....28
2.4.2.2 Reconocimiento de palabras aisladas....30
2.4.2.3 Reconocimiento de habla continua....42
3. RECONOCEDOR AUTOMÁTICO DE HABLA EN CASTELLANO....47
3.1 Resumen....47
3.2 Reconocedor automático en inglés....48
3.2.1 Base de datos....49
3.2.2 Diccionario....50
3.2.3 Tipo de parametrización....52
3.2.4 Topología de los modelos utilizados....52
3.2.5 Fases del reconocedor....54
3.2.6 Experimentos realizados....56
3.3 Reconocedor automático en castellano....57
3.3.1 Base de datos....58
3.3.2 Diccionario....60
3.3.3 Tipo de parametrización....64
3.3.4 Topología de los modelos utilizados....65
3.3.5 Fases del reconocedor....67
3.3.6 Experimentos realizados....77
4. EFECTOS DEL GÉNERO DEL LOCUTOR SOBRE EL RECONOCIMIENTO DE HABLA....79
4.1 Resumen....79
4.2 Adaptación MAP....80
4.3 Clasificador de género....83
4.3.1 Principios de desarrollo....84
4.3.2 Estructura del clasificador de género....87
4.3.3 Experimentos de clasificación de género....89
4.3.3.1 Datos independientes del locutor....90
4.3.3.2 Datos dependientes del locutor....94
4.3.4 Conclusiones....97
4.4 Experimentos realizados....98
4.4.1 Adaptación MAP al género....98
4.4.1.1 Utilización del género real en la fase de entrenamiento....100
4.4.1.2 Utilización del género decidido por el clasificador en la fase de entrenamiento 103
4.4.2 Entrenamiento completo por género....105
4.5 Conclusiones....107

5. EFECTOS DEL LOCUTOR SOBRE EL RECONOCIMIENTO DE HABLA 109
5.1 Resumen 109
5.2 Adaptación al locutor 110
5.2.1 Adaptación MAP 111
5.2.2 Adaptación MLLR 111
5.3 Experimentos realizados 113
5.3.1 Adaptación MAP al locutor 113
5.3.2 Adaptación MLLR al locutor 116
5.3.2.1 Adaptación MLLR incremental 117
5.3.2.2 Adaptación MLLR supervisada 120
5.3.3 Entrenamiento completo por locutor 121
5.3.4 Combinación de adaptación al género y al locutor 122
5.3.4.1 Adaptación MAP al género y adaptación MAP al locutor 123
5.3.4.2 Adaptación MAP al género y adaptación MLLR al locutor 125
5.4 Conclusiones 127
6. CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS 129
6.1 Conclusiones 129
6.2 Líneas de trabajo futuras 133
7. PRESUPUESTO 135
7.1 Introducción 135
7.2 Coste del material 136
7.3 Coste del personal 137
7.4 Presupuesto total 138
8. REFERENCIAS 139
9. ANEXO I 143

Índice de figuras

Figura 1.- Aparato fonador humano (obtenida a partir de [REV]) 11
Figura 2.- Componentes de las cavidades supraglóticas (obtenida a partir de [REV]) ... 13
Figura 3.- Sistema auditivo periférico (obtenido a partir de [EUMUS]) 15
Figura 4.- Diagrama de bloques de generación de habla 16
Figura 5.- Diagrama de bloques de recepción de habla 17
Figura 6.- Esquema del reconocimiento de un mensaje 18
Figura 7.- Etapas de un sistema de reconocimiento automático de habla 19
Figura 8.- Enventanado necesario para la parametrización 21
Figura 9.-Banco de filtros equiespaciados en escala Mel y el vector de parámetros correspondiente, siendo N el número de filtros 25
Figura 10.- Ejemplo de modelo de Markov 32
Figura 11.- Mezcla de M gausianas 35
Figura 12.- Algoritmo de Viterbi 41
Figura 13.- Topología HMMs fonemas 53
Figura 14.- Topología de HMM 'sil' 53
Figura 15.- Topología de HMM 'sp' 53
Figura 16.- modificación de trascripción al pasar a trifonemas 71
Figura 17.- Ejemplo de modelo de lenguaje equiprobable 74
Figura 18.- Ejemplo de salida de HResult 76
Figura 19.- Topología de los GMMs del clasificador de género 88

Figura 20.- Clasificador de género 89
Figura 21.- Grupos independientes del locutor – 1 reestimación 91
Figura 22.- Grupos independientes del locutor – 2 reestimaciones.....92
Figura 23.- Grupos independientes del locutor – 3 reestimaciones.....93
Figura 24.- Grupos independientes del locutor – ajuste del umbral 94
Figura 25.- Grupos dependientes del locutor – 3 reestimaciones 95
Figura 26.- Grupos dependientes del locutor – ajuste del umbral 96
Figura 27.- Entrenamiento del reconocedor de habla con adaptación al género 99
Figura 28.-Reconocimiento con modelos adaptados al género 100
Figura 29.- Barrido de τ (adaptación con género real, reconocimiento con género del clasificador) 103
Figura 30.- Barrido de τ (adaptación y reconocimiento con género del clasificador) 105
Figura 31.- Entrenamiento del reconocedor de habla con modelos separados por género 106
Figura 32.- Entrenamiento del reconocedor de habla con adaptación MAP al locutor .114
Figura 33.- Reconocimiento con modelos adaptados al locutor.....115
Figura 34.- Adaptación MAP al locutor (barrido de τ).....116
Figura 35.- Fase de adaptación para MLLR incremental no supervisada 118
Figura 36.- Fase de reconocimiento para adaptación MLLR 118
Figura 37.-Adaptación al locutor de modelos adaptados al género 122
Figura 38.- Adaptación MAP al género y al locutor.....124
Figura 39.- Reconocimiento para locución del locutor ‘N’ con modelos adaptados al género con MAP y al locutor con MLLR.....126

Índice de tablas

Tabla 1.- Conjunto de fonemas utilizados 51
Tabla 2.- Información locutores en función de variedad dialectal y género. 58
Tabla 3.- Información locutores en función de edad y género. 59
Tabla 4.- Conjunto de fonemas utilizados 61
Tabla 5.- conjunto de unidades acústicas final 62
Tabla 6.- cambio de caracteres no permitidos por HTK. 62
Tabla 7.- cambio de caracteres no permitidos por SAGA 63
Tabla 8.- Costes de material 137
Tabla 9.- Costes de personal 138
Tabla 10.- Clasificador de género, grupos independientes del locutor, 1 reestimación 143
Tabla 11.- Clasificador de género, grupos independientes del locutor, 2 reestimaciones 143
Tabla 12.- Clasificador de género, grupos independientes del locutor, 3 reestimaciones 144
Tabla 13.- Clasificador de género, grupos independientes del locutor, ajuste del umbral 144
Tabla 14.-Clasificador de género, grupos dependientes del locutor,3 reestimaciones . 145
Tabla 15.- Clasificador de género, grupos dependientes del locutor, ajuste del umbral 145

Tabla 16.- Barrido de τ, utilizando género real en la fase de entrenamiento y género del clasificador en reconocimiento. Se adaptan medias y varianzas.....145
Tabla 17.- Barrido de τ, utilizando género del clasificador en la fase de entrenamiento y género del clasificador en reconocimiento. Sólo se adaptan las medias.....146
Tabla 18.- Barrido de τ, realizando adaptación MAP al locutor. Sólo se adaptan las medias.....146
Tabla 19.- Barrido de τ, realizando adaptación MAP al locutor y adaptación MAP al género. Sólo se adaptan las medias.....148

Capítulo 1

Introducción y objetivos

1.1 Introducción

El presente proyecto fin de carrera trata sobre el análisis de técnicas que, mediante el tratamiento de la información del género o el locutor, permitan mejorar las prestaciones de un reconocedor automático de habla continua en castellano.

Para esto, y tras el desarrollo del reconocedor inicial, se han realizado las modificaciones pertinentes para la adaptación del sistema a distintas situaciones planteadas en función de dicha información.

En primer lugar se ha utilizado la técnica de adaptación MAP para adaptar los modelos del reconocedor inicial a la información del género, lo que genera dos nuevos conjuntos de modelos, cada uno de ellos representativo de un género diferente. En este caso, en la fase de evaluación del reconocedor, se eligen los modelos correspondientes al género de la locución para realizar el reconocimiento.

Además de lo anterior, se ha evaluado la posibilidad de crear los modelos dependientes del género sin necesidad de realizar adaptación, con un entrenamiento completo con los datos divididos según el género.

Para permitir una detección automática del género de las locuciones se ha desarrollado un clasificador de género, que puede aplicarse tanto en la fase de entrenamiento como en la de test del reconocedor.

Una vez realizado el estudio anterior, se ha procedido a evaluar la influencia de la información sobre el locutor, adaptando los modelos iniciales con los datos correspondientes a cada uno de ellos. Se han utilizado dos técnicas diferentes: MAP y MLLR, lo que además de permitir conocer si el uso de esta información mejora los resultados de reconocimiento, proporciona información sobre las prestaciones de ambas técnicas. En este caso, en la fase de test, el reconocedor elige el conjunto de modelos adecuado al locutor bajo test, ya que dispone de uno diferente para cada locutor sobre el que se realiza la adaptación.

Una vez estudiado el comportamiento del sistema ante la información del género y del locutor, se ha planteado la utilización combinada de ambas informaciones, combinando también las técnicas de adaptación utilizadas previamente.

1.2 Objetivos

El objetivo principal del proyecto es mejorar los resultados de un reconocedor automático de habla continua, utilizando para ello información sobre el género y sobre el locutor.

Para ello se va a modificar el reconocedor inicial, provocando que la generación de los modelos de las distintas unidades acústicas utilizadas dependa de dicha información. En función de los datos utilizados en el proceso se pretende deducir si la señal de voz modela características representativas del género o del locutor, y cuáles de ellas proporcionan un mayor beneficio en el reconocimiento, evaluándose:

- El uso aislado de la información sobre el género.
- El uso aislado de la información sobre el locutor.
- Una combinación de la información sobre el género y sobre el locutor.

En el caso concreto de la adaptación al género, el objetivo es evaluar la utilización de una técnica de adaptación (MAP) o el entrenamiento completo de los modelos representativos de

cada género. Con esto, además de mejorar las prestaciones del reconocedor, se pretende evaluar qué técnica ofrece mejores resultados, deduciendo si los datos son suficientemente extensos como para no necesitar técnicas de adaptación.

Además, para la elección del género puede utilizarse un clasificador automático de género. La inclusión del mismo en el proceso de reconocimiento tiene como objetivo mejorar los resultados respecto al uso del género real, ya que agrupará las muestras de voz en función del género con el que comparte más características, siendo éstas las que pretenden modelarse a partir de dicha agrupación.

Para la adaptación al locutor, el objetivo es evaluar dos técnicas de adaptación diferentes, MAP y MLLR, para determinar la que aporta mayores beneficios al reconocimiento.

Una vez se conozca la inclusión de qué información proporciona mejores beneficios, y qué técnica debe usarse para ello, se combinará la información del género y del locutor, combinando también las técnicas de adaptación, con el objetivo de deducir si merece la pena el aumento de complejidad producido respecto a la mejora obtenida.

1.3 Esquema

En el presente apartado se ofrece una breve explicación de la estructura utilizada en esta memoria, detallando el contenido de los distintos capítulos que la componen.

Se ha decidido utilizar un total de 6 capítulos, cuyo contenido particular se muestra a continuación:

- Capítulo 1: ofrece una introducción del desarrollo llevado a cabo en el proyecto fin de carrera, enmarcado en el ámbito de las tecnologías del habla. Tras enumerar los objetivos que se persiguen, se realiza una breve presentación de dichas tecnologías, ofreciéndose una idea sobre su campo de actuación.
- Capítulo 2: expone las bases teóricas del reconocimiento de habla.
- Capítulo 3: aplica las bases teóricas expuestas en el Capítulo 2 al desarrollo de un reconocedor automático de habla continua en castellano.
- Capítulo 4: explica las modificaciones realizadas sobre el reconocedor original para que se tenga en cuenta la información de género, y las distintas pruebas realizadas para su evaluación. También se describe el desarrollo del clasificador de género y su utilización en la evaluación mencionada.

- Capítulo 5: describe la inclusión de la información sobre el locutor en el proceso de reconocimiento, y las prestaciones que se obtienen. Además, contiene la descripción de los desarrollos realizados para combinar tanto la información del género como la del locutor, y cómo afecta esto a los resultados del reconocedor.
- Capítulo 6: en este capítulo se resumen las conclusiones obtenidas a lo largo del proyecto, basadas en los resultados experimentales mostrados a lo largo de los Capítulos 3, 4 y 5. Por último contiene una serie de reflexiones realizadas sobre posibles líneas de actuación futuras.

1.4 Tecnologías del habla

Este proyecto fin de carrera se centra en el reconocimiento automático de habla, siendo ésta una tecnología perteneciente a las conocidas como tecnologías del habla.

Tecnologías del habla es un término que agrupa una variedad de disciplinas que surgen del estudio del habla, con el objetivo de reproducir artificialmente los procesos implicados en la comunicación oral [Fer03].

1.4.1 Clasificación

A continuación se presenta una clasificación de las tecnologías del habla en función de las tecnologías básicas que las componen [Her03]:

- *Codificación de voz:* comprende el tratamiento de la señal de voz, obteniendo sus características más representativas. Los parámetros obtenidos representan los sonidos de forma más efectiva para su posterior tratamiento, lo que resulta fundamental para el reconocimiento y síntesis de habla, y además reduce el volumen necesario para el almacenamiento de estos datos [CEIDIS].
- *Síntesis de habla:* se trata de la generación automática de una señal de voz, pudiéndose diferenciar entre sistemas de síntesis que se limitan a un vocabulario restringido y sistemas que pueden llegar a transformar cualquier texto escrito, en formato electrónico, en su representación acústica [Lli09], [CEIDIS].
- *Reconocimiento automático de habla:* su objetivo es transformar la señal de voz en una representación de la misma, normalmente en formato escrito. Se pueden clasificar en función del vocabulario soportado, del tipo de locución de entrada

(desde palabras aisladas hasta habla continua), o incluso en función de si el sistema está adaptado o no al locutor [Lli09]. Es en el marco de esta tecnología en el que se va a trabajar a lo largo del presente proyecto fin de carrera.

Analizando la evolución histórica reciente de las tres tecnologías anteriores, se concluye que todas ellas cumplen un modelo evolutivo común: partiendo de un modelo basado en el conocimiento de la teoría general de la lengua, en un determinado momento evolucionan hacia un modelado gobernado por los datos. Es en este punto donde empiezan a obtenerse mejores resultados y donde la potencia de cálculo empieza a tomar un papel fundamental [Her03], [ML96].

1.4.2 Evolución histórica

En este apartado se pretende realizar un resumen de la evolución histórica de las tecnologías del habla [Fer03],[Her03].

En cuanto a la síntesis de habla los primeros desarrollos se sitúan en 1791, utilizando modelos mecánicos. Mediante un fuelle se generaba una presión que atravesaba unas lengüetas, haciendo éstas las veces de cuerdas vocales, y finalizaba en unos tubos que, en función de la deformación aplicada a los mismos, producían diferentes sonidos.

En 1972 el modelado se basa ya en conocimientos de tratamiento de la señal y de procesamiento digital. En este caso, se dispone de un esquema de filtros que son atravesados por señales de excitación y que generan la primera voz sintética con un nivel de inteligibilidad aceptable.

A partir de este momento, los desarrollos van dirigidos a modificar los parámetros del sistema en función del tiempo, para aumentar la calidad y las características de la voz, lo que implica un análisis muy elevado de texto de entrada para obtener la información que hay que manejar.

En cuanto al reconocimiento de habla, los primeros trabajos estaban muy relacionados con los realizados en síntesis de habla, ya que partiendo de los modelos de síntesis se podía conocer la información más relevante de la señal de voz, centrándose en dichas características para realizar el análisis propio del reconocimiento.

La evolución en reconocimiento de habla comienza con los reconocedores cuasifonéticos (1950). Éstos utilizaban diccionarios muy reducidos y trabajaban a partir de medidas de energía,

cruces por cero o modelos LPC de bajo orden, marcando reglas simples de estas medidas que describían los distintos alófonos de las palabras que formaban el diccionario.

A continuación aparecen las técnicas de alineamiento dinámico temporal, necesarias debido al hecho de que repeticiones de una misma locución presentan temporizaciones muy diversas de cada uno de los eventos acústicos que la componen.

A finales de los años 60, principios de los 70, aparece el reconocimiento de patrones, entre los que las redes neuronales y los modelos de Markov han sido los más extendidos. Se trata de técnicas estadísticas de clasificación, que necesitan de una gran cantidad de datos de entrenamiento.

A partir de los años 80 se fue añadiendo el conocimiento del lenguaje natural a las tecnologías del habla, surgiendo una serie de tecnologías de integración que combinan las clasificadas anteriormente como tecnologías básicas, siendo la más representativa el control de diálogo (sistemas de voz interactivos).

Es necesario mencionar, que el avance tecnológico conseguido en las tecnologías del habla ha sido posible gracias al aumento en capacidad de cálculo y memoria de los recursos tecnológicos utilizados.

La situación actual se puede resumir como sigue:

- *Conversores texto-voz*:
  - Inteligibilidad excelente.
  - Problemas de naturalidad.
  - Primeros avances en reproducción de emociones.
  - Primeros pasos en la generación de nuevas voces sin necesidad de gran cantidad de datos de entrenamiento.
- *Reconocimiento de habla*:
  - Tasas altas de reconocimiento para vocabularios medios.
  - Problemas de robustez en función de las condiciones de trabajo.
  - Problemas con el habla espontánea.
- *Control de diálogo*:
  - Soluciones para dinamizar y adaptar la secuencia de intervenciones.
  - Funcionan bien en entornos restringidos, como el caso concreto de recuperación de información.
  - Limitaciones en los sistemas de síntesis y reconocimiento de habla utilizados.
  - Dificultad de desarrollo por la búsqueda de una comunicación hombre-máquina lo más parecida posible a la comunicación entre personas.

1.4.3 Líneas futuras de desarrollo

La evolución de las tecnologías del habla [Fer03][ML96][Her03] debe solucionar los problemas presentes en la actualidad, planteándose también nuevas líneas de desarrollo. A continuación se enumeran los posibles y deseables puntos hacia los que parecen tender dichas tecnologías:

- *Conversores texto-voz:*
  - Mejorar la naturalidad, haciendo el sistema más agradable al oyente.
  - Desarrollo de voces nuevas de forma automática y con pocas muestras.
  - Expresión de estados de ánimo y emociones.
  - Aumento de la aceptación del usuario.
- *Reconocimiento de habla:*
  - Mejoras en reconocimiento de habla espontánea, con aprendizaje dinámico de expresiones.
  - Robustez ante fuentes de variabilidad.
- *Control de diálogo:*
  - Mejora del flujo de diálogo.
  - Dinamicidad de las intervenciones.
  - Modularidad de diálogos, para hacerlos reutilizables.
  - Aprendizaje automático.
  - Mayor facilidad de uso, con un aumento del número de palabras utilizadas y mejoras en la naturalidad del sistema, lo que podrá conseguirse mediante mejoras en las tecnologías de los puntos anteriores.

Paralelo a todo esto debe existir una evolución en herramientas para el etiquetado automático de habla, haciendo posible el desarrollo de aquellos sistemas que necesiten gran cantidad de datos de entrenamiento.

Como línea de desarrollo más exigente, se puede pensar en sistemas de diálogo sin funcionalidad específica, con los que se pueda dialogar para finalidades muy diferentes. Dichos sistemas deberían poder extraer la intención con la que se está desarrollando la comunicación, de manera que proporcionen justo lo que necesita el usuario de ese sistema. Además deberían ser sistemas con capacidad continua de autoaprendizaje, robustos ante variaciones del canal de comunicación, del locutor o ante expresiones de habla espontánea. También parece interesante trabajar sobre el interfaz que presentan estos sistemas para complementar aquellas funcionalidades de difícil control por voz, siendo deseable el uso de interfaces gráficas, textos e incluso información gestual.

Otro aspecto importante es el concepto de multilingualidad, que implica el desarrollo de sistemas de reconocimiento y síntesis de habla multilingües, así como el de sistemas de traducción automática voz-voz y de sistemas de gestión de diálogo independientes del idioma de los miembros que intervengan en la comunicación.

Como punto final mencionar que el motivo principal por el que la comunicación hombre-máquina no consigue igualarse a la comunicación entre personas es porque aún no se tiene un conocimiento completo de las funciones realizadas por el cerebro en este proceso. Es fundamental descubrir la estructura fundamental de su funcionamiento para alcanzar este reto.

Capítulo 2

Reconocimiento automático de habla

2.1 Resumen

El presente capítulo pretende realizar una revisión sobre los conceptos teóricos necesarios para el desarrollo de este proyecto fin de carrera.

En primer lugar resulta fundamental obtener información acerca de los procesos que intervienen en la comunicación oral humana, ya que serán estos procesos los que se intentan simular de manera automática en todo sistema perteneciente a las tecnologías del habla. En particular, se va a estudiar el comportamiento ante el reconocimiento de habla, que es el objetivo del presente proyecto fin de carrera.

Una vez presentado el comportamiento del ser humano ante el problema del reconocimiento de habla, se estudiará la forma de simular este comportamiento para realizar un sistema de

reconocimiento automático de habla. Se expondrán las distintas etapas que forman dicho sistema, y las opciones de desarrollo que presentan cada una de ellas, prestando especial atención a las opciones elegidas para el reconocedor de habla desarrollado.

2.2 Introducción

En este apartado se estudian las distintas fases en las que se puede dividir la comunicación oral entre dos personas [HAH01].

Toda comunicación oral comienza con la intención de comunicar algún tipo de información por parte del locutor. Su cerebro será el encargado de generar un mensaje, dotándole del significado esperado. El mensaje será dividido en las distintas unidades acústicas, responsables de ofrecer información sobre la pronunciación del mensaje. Además, el sistema puede introducir variaciones en cuanto a la duración de cada fonema o cambios en la entonación. La información anterior se transforma en información neuromuscular con la que se controlan los distintos elementos del tracto vocal, de tal forma que se emitan los sonidos deseados.

Una vez la señal de voz es emitida, ésta se transmite y llega al receptor, que trabajará en orden inverso al locutor. En este caso el punto de entrada es el sistema auditivo, que se encarga de tratar la señal de entrada y prepararla para el resto del sistema. Una vez tratada, se transforma en información neuronal entendible por parte del cerebro, que ayudado de información sobre el sistema de lenguaje y semántica, obtendrá el mensaje que se transmitió.

Como ya se ha comentado, las tecnologías del habla pretenden simular este proceso de forma automática.

2.3 Producción y recepción de voz

En todos los ámbitos que engloban las tecnologías del habla es necesario conocer cómo el ser humano produce la señal de voz. Sólo partiendo de este conocimiento se puede estudiar la onda de sonido generada y extraer cuáles serán sus características más representativas, fundamental no sólo para la codificación, sino también para el reconocimiento o síntesis de voz.

El objetivo de este apartado es ofrecer una visión general sobre los procesos que forman parte de la generación y recepción de la voz, ofreciendo el enfoque inicial necesario para afrontar un tratamiento computacional de estos procesos.

2.3.1 Generación de la señal de voz

La voz no es más que sonido emitido por las personas. Sin embargo se trata de un sonido dotado de una serie de características especiales, que son aportadas por el aparato fonador.

El sonido es una onda longitudinal de presión, y su información estará contenida en las variaciones de presión presentes en la misma. En el caso de la voz, será el aparato fonador el encargado de modular dicha presión para transmitir la información deseada.

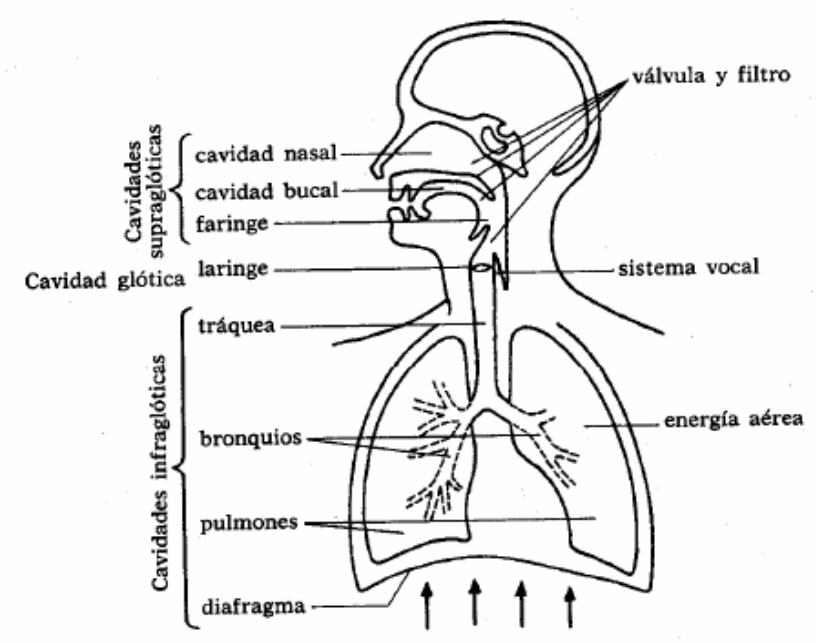

Figura 1.- Aparato fonador humano (obtenida a partir de [REV])

En la Figura 1 se muestra un diagrama de los principales componentes del aparato fonador humano. En él se observan tres partes bien diferenciadas [Miy04]:

- *Cavidades infraglóticas*: formadas por el diafragma, los pulmones, los bronquios y la traquea.
- *Cavidad glótica*: formada por la laringe, que es la cavidad donde se encuentran, entre otros elementos, las cuerdas vocales. Se trata de 4 músculos, de los cuales

sólo 2 intervienen en el proceso de generación de voz. El hueco que dejan estos dos músculos es lo que se conoce como glotis.

- *Cavidades supraglóticas*: formadas por la faringe, la cavidad bucal y la cavidad nasal.

El proceso de generación de voz [FUR89], [RJ93], [HAH01], comienza en las cavidades infraglóticas, que son las que aportarán la fuente de aire necesaria. El aire que se encuentra en los pulmones será expulsado por estos con la ayuda del diafragma, y gracias a los bronquios y a la traquea éste aire llegará a la cavidad glótica. Una vez allí, el aire se encuentra con la glotis, y al chocar contra ella se generará el sonido, produciéndose un sonido 'sonoro' si el choque del aire provoca que las cuerdas vocales vibren, o bien un sonido 'sordo' si no se produce vibración. Aunque en este punto ya se habla de sonido, es en las cavidades supraglóticas donde se amplifica y se modula para generar toda la variedad de sonidos que forman la voz. Tanto la faringe, como la cavidad nasal y bucal actúan como cajas de resonancia, y los distintos elementos que las componen (ya sean móviles o estáticos) actúan como articuladores, permitiendo cambiar algunas características para adaptar esos sonidos al habla.

El sonido así generado es la realización física de un fonema, que no es más que la imagen mental de un determinado sonido. Mientras que los fonemas que forman una lengua son un conjunto limitado, los sonidos producidos al hablar pueden ser infinitos. Incluso una misma persona al pronunciar varias veces una misma palabra puede introducir variaciones (que deberán aislarse en el reconocimiento de habla, puesto que no aportan información lingüística).

Una primera clasificación de los fonemas [Miy04] sería la división en vocales y consonantes:

- *Vocales*: todos ellos son sonidos 'sonoros', es decir, producidos por la vibración de las cuerdas vocales. En ellos la onda de aire no encuentra obstáculo alguno al atravesar las cavidades supraglóticas para salir al exterior.
- *Consonantes*: en este caso existen obstáculos en el recorrido del aire hasta el exterior.

Se puede hacer una clasificación mucho más exhaustiva de los fonemas que pertenecen a cada grupo, teniendo en cuenta el punto de articulación, el modo de articulación o la intervención de la cavidad nasal.

En el caso del castellano, las vocales se pueden clasificar como:

- *Anterior, medio o central*, en función del lugar de la cavidad bucal donde se realiza la modulación de ese fonema (punto de articulación).
- *De abertura máxima, media o semiabierta*, en función de la abertura de la boca (modo de articulación).

Todas las vocales del castellano son sonoras y en ellas no interviene la cavidad nasal (el velo del paladar impide el paso del aire hacia esta cavidad).

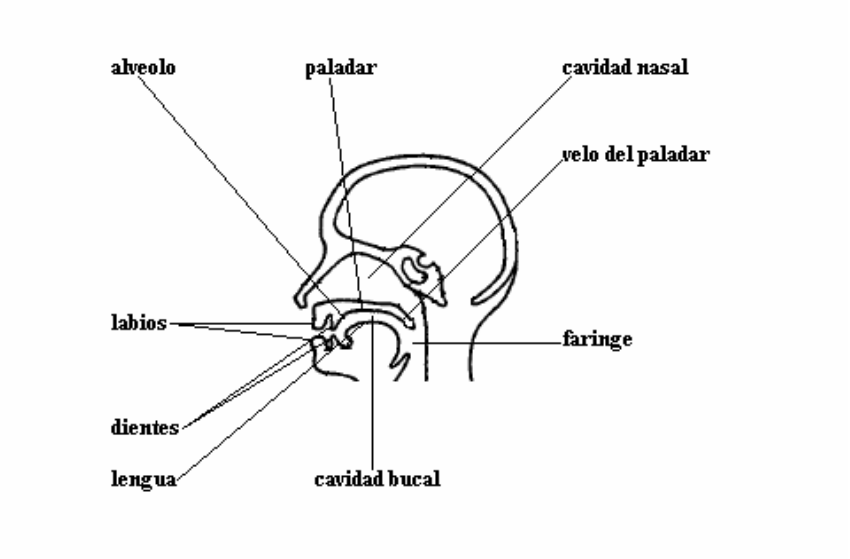

Figura 2.- Componentes de las cavidades supraglóticas (obtenida a partir de [REV])

En el caso de las consonantes del castellano, algunas clasificaciones posibles son:

- En función del *punto de articulación*, refiriéndose al lugar del aparato fonador (Figura 2) donde se produce la obstrucción al paso del aire para las consonantes:
  - Bilabial: unión de los labios.
  - Labiodental: labio inferior y dientes superiores.
  - Linguodental: lengua y dientes superiores.
  - Alveolar: lengua y alvéolos superiores.
  - Palatal: lengua y paladar.
  - Velar: lengua y velo del paladar.
  - Glotal: articulación en la glotis.
- En función del *modo de articulación*, o postura que adoptan los distintos órganos que intervienen:
  - Fricativo: se produce un estrechamiento, de forma que el aire pasa rozando.
  - Oclusivo: cierre total, pero momentáneo, al paso del aire.
  - Africado: oclusión seguida de fricación.
  - Lateral: el aire sale al exterior rozando los lados de la cavidad bucal.
  - Vibrante: vibra la punta de la lengua al pasar el aire produciendo obstrucción intermitente.
  - Aproximante: obstrucción en una zona muy estrecha, de forma que no produce turbulencia.
- Las consonantes también se pueden agrupar en sonidos 'sonoros' y 'sordos' en función de la vibración o no vibración de las cuerdas vocales.

- Además, en la generación de las consonantes puede intervenir, o no, la cavidad nasal (el velo del paladar puede cerrar el paso a esta cavidad), tratándose de sonidos 'nasales' u 'orales'.

Lo realmente importante desde el punto de vista del tratamiento de la señal de voz es conocer cómo afecta todo el proceso anterior en las características de dicha señal[HAH01].

La principal diferencia entre los sonidos sordos y sonoros es que estos últimos presentan una estructura mucho más regular que los primeros (tanto en su estructura temporal como frecuencia), al igual que una mayor intensidad, lo que se debe principalmente a la vibración de las cuerdas vocales en la generación de los sonidos sonoros, provocando que muestren una estructura regular. La frecuencia con la que vibran las cuerdas vocales es conocida como frecuencia fundamental o frecuencia de pitch, y su valor puede variar desde 60Hz hasta más de 300Hz en función del locutor.

Una vez el aire ha pasado las cuerdas vocales, el resto del sistema fonador se encarga de modificar su estructura frecuencial. Si se tiene un sonido sonoro, es decir, una onda de presión con una frecuencia fundamental y sus armónicos, las cavidades resonantes con las que se encuentre modificarán esta estructura, realizándose una amplificación de las frecuencias cercanas a las frecuencias de resonancia, conocidas como formantes. En función del fonema a pronunciar, la forma de las cavidades se verá modificada, y por lo tanto también los formantes generados.

La voz se puede considerar como una concatenación de los sonidos descritos, por lo que las características del sistema irán variando en el tiempo, en función de los sonidos a producir para generar el mensaje [RJ93].

2.3.2 Recepción de la señal de voz

El proceso de recepción de la voz es llevado a cabo por el sistema auditivo humano. Este sistema se puede dividir en dos partes [HAH01]: el sistema auditivo periférico, que permite la recepción de los sonidos y envía la información correspondiente al cerebro, y el sistema auditivo central, que proporciona un significado a estos sonidos.

El funcionamiento del sistema auditivo central se escapa del ámbito de este proyecto, por lo que únicamente se tratará en este apartado el sistema auditivo periférico.

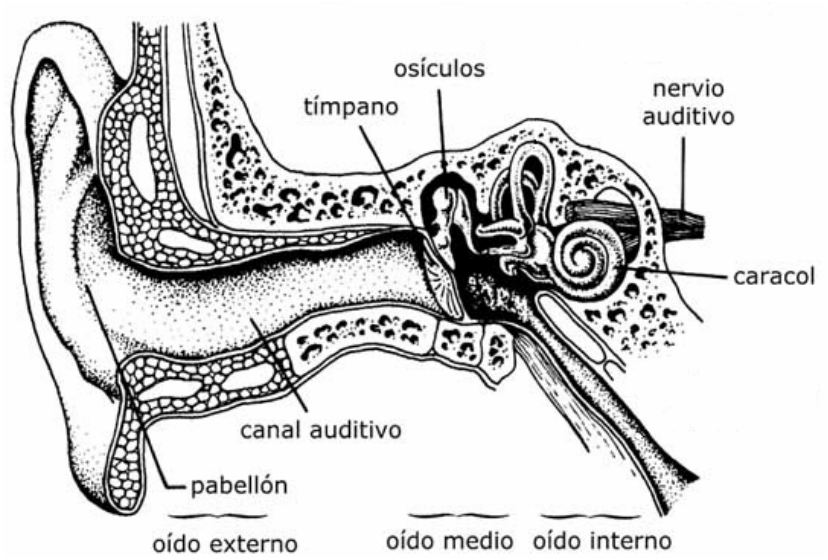

Figura 3.- Sistema auditivo periférico (obtenido a partir de [EUMUS])

En la Figura 3 se muestra un diagrama de los principales componentes del sistema auditivo periférico humano. En él se observan tres partes bien diferenciadas:

- *Oído externo*: formado por la parte externa y el canal auditivo externo.
- *Oído medio*: se trata de una cavidad rellena de aire donde se encuentran el tímpano y los osículos.
- *Oído interno*: su elemento principal es la cóclea o caracol.

El proceso de recepción auditiva comienza en el oído externo, encargado de transportar la onda de presión acústica hasta el tímpano (que separa el oído externo del oído medio). Debido a las variaciones de presión que recibe, el tímpano vibra, y esta vibración se transmite al lado contrario del tímpano, a los huesos conocidos como osículos, con una frecuencia igual a la del sonido recibido. Por último, esta vibración es transmitida al oído interno mediante la ventana oval, que es una membrana que hace de interfaz entre el oído medio y el interno. Ya en la cóclea, la vibración se transforma en impulsos eléctricos que se transmitirán al cerebro a través del nervio auditivo.

Todo el proceso llevado a cabo en el sistema auditivo provoca diferentes efectos en la recepción de sonidos [FUR89], [HAH01], como pueden ser:

- *Variación de la sensibilidad del oído*: la sensibilidad del oído depende de la frecuencia y de la calidad del sonido. Debido a las resonancias que presenta el canal del oído externo, se produce un aumento de la sensibilidad justo en las frecuencias de dichas resonancias (en torno a los 4 KHz y los 13 KHz).
- *Enmascaramiento*: si tenemos dos tonos muy cercanos, ya sea en tiempo o en frecuencia, el oído no puede diferenciar esos tonos. A mayor intensidad de un tono, mayor será el rango de enmascaramiento que produce.

El oído realiza un tratamiento frecuencial sobre la voz de entrada, de forma que extrae las características necesarias para reconocer el fonema, con los problemas ya mencionados. Estas características serán enviadas hacia el cerebro mediante impulsos eléctricos.

2.3.3 Enfoque computacional

En los apartados 2.3.1 y 2.3.2 se ha ofrecido una visión general sobre los mecanismos y características que envuelven la generación y la recepción de habla, conocimiento imprescindible cuando se trabaja en el ámbito de las tecnologías del habla.

Llegado este punto es interesante conocer como modelar los procesos anteriores desde un punto de vista computacional. En el presente apartado se aportan los diagramas de bloques correspondientes, entendiéndose estos como puntos de partida de cualquier tratamiento a realizar.

En cuanto a la generación de voz se ha visto que el aparato fonador genera una señal periódica o ruidosa, en función de si se va a producir un sonido sonoro o sordo, que se verá modulada por una serie de cavidades para generar el sonido deseado. En la Figura 4 se puede ver un diagrama de bloques que representa este sistema.

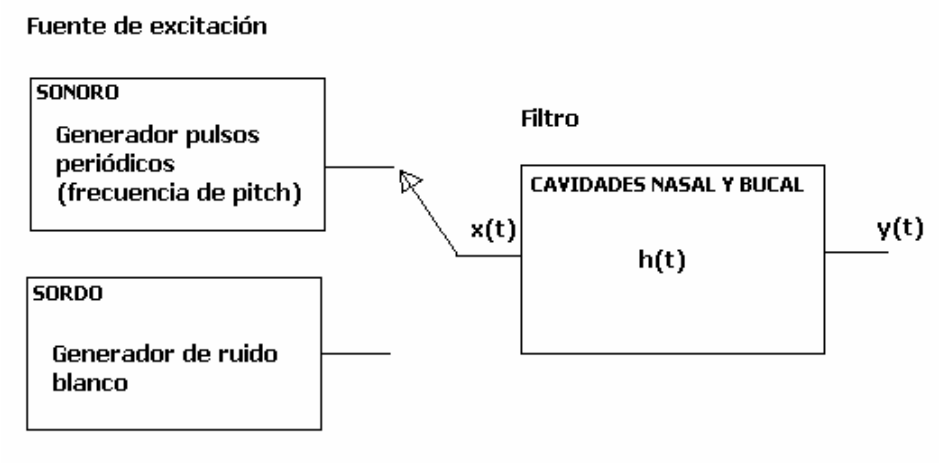

Figura 4.- Diagrama de bloques de generación de habla

En cuanto a la recepción del habla se ha comentado que la respuesta frecuencial del oído no es lineal. Se puede simular su comportamiento como un analizador espectral del sonido. Se conoce una serie de bandas críticas en la respuesta de la cóclea, lo que hace que se pueda simular por una sucesión de filtros solapados, con anchura igual a la de esas bandas críticas.

Una vez realizado el análisis espectral, el cerebro será el encargado de decidir a qué fonemas se corresponde la información resultante, y obtener por sucesión de estos patrones las palabras o frases pronunciadas. En la Figura 5 se puede ver un diagrama de bloques de este proceso.

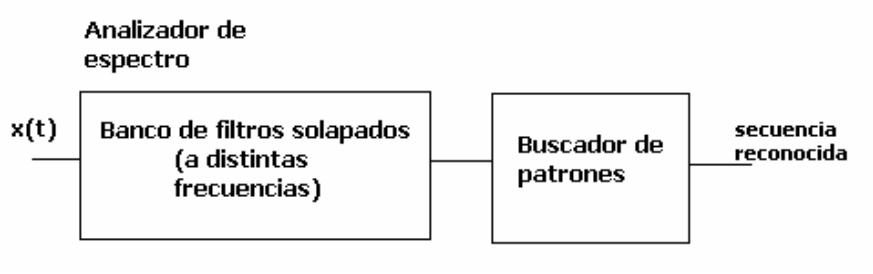

Figura 5.- Diagrama de bloques de recepción de habla

2.4 Reconocimiento automático de habla

Un sistema de reconocimiento automático de habla es aquel que partiendo de la señal de voz es capaz de obtener el mensaje que se está transmitiendo.

Como ya se ha comentado, el proceso de generación de habla comienza cuando el locutor se hace una idea mental de lo que quiere comunicar para, a continuación, emitir la señal de voz correspondiente a esa idea inicial. Por lo tanto, un sistema de reconocimiento automático de habla debe intentar obtener esa información inicial partiendo de la señal de voz que le está llegando. En la Figura 6 queda representado este proceso: la señal de voz será tratada para obtener sus características más representativas (parametrización), de forma que permita recuperar el mensaje inicial (representado en la Figura 6 como la sucesión de símbolos S1, S2 y S3).

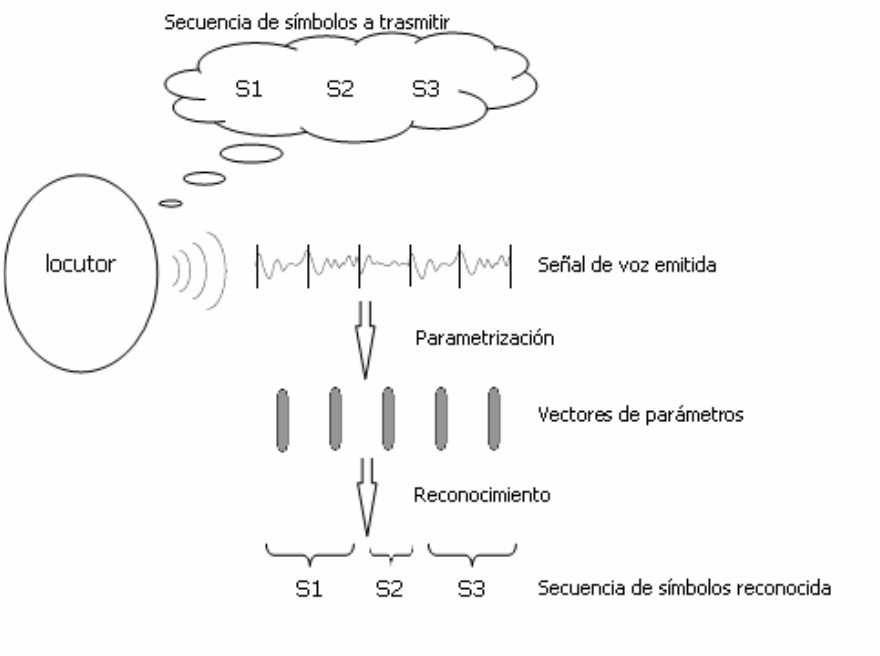

Figura 6.- Esquema del reconocimiento de un mensaje

Para este proyecto en particular, la salida del reconocedor será una trascripción del mensaje de cada locución de entrada.

El proceso de reconocimiento automático de habla se puede entender como un sistema de reconocimiento de patrones [HAH01], [RJ93]. Esta afirmación se basa en que si el sistema es capaz de, partiendo de la señal de voz, obtener la información adecuada para representar dicha señal con una sucesión de modelos, bastaría con comparar esos patrones con unos ya conocidos para decidir el mensaje que le está llegando. En la Figura 7 se muestra un diagrama de bloques de las distintas etapas de un sistema de reconocimiento automático de habla.

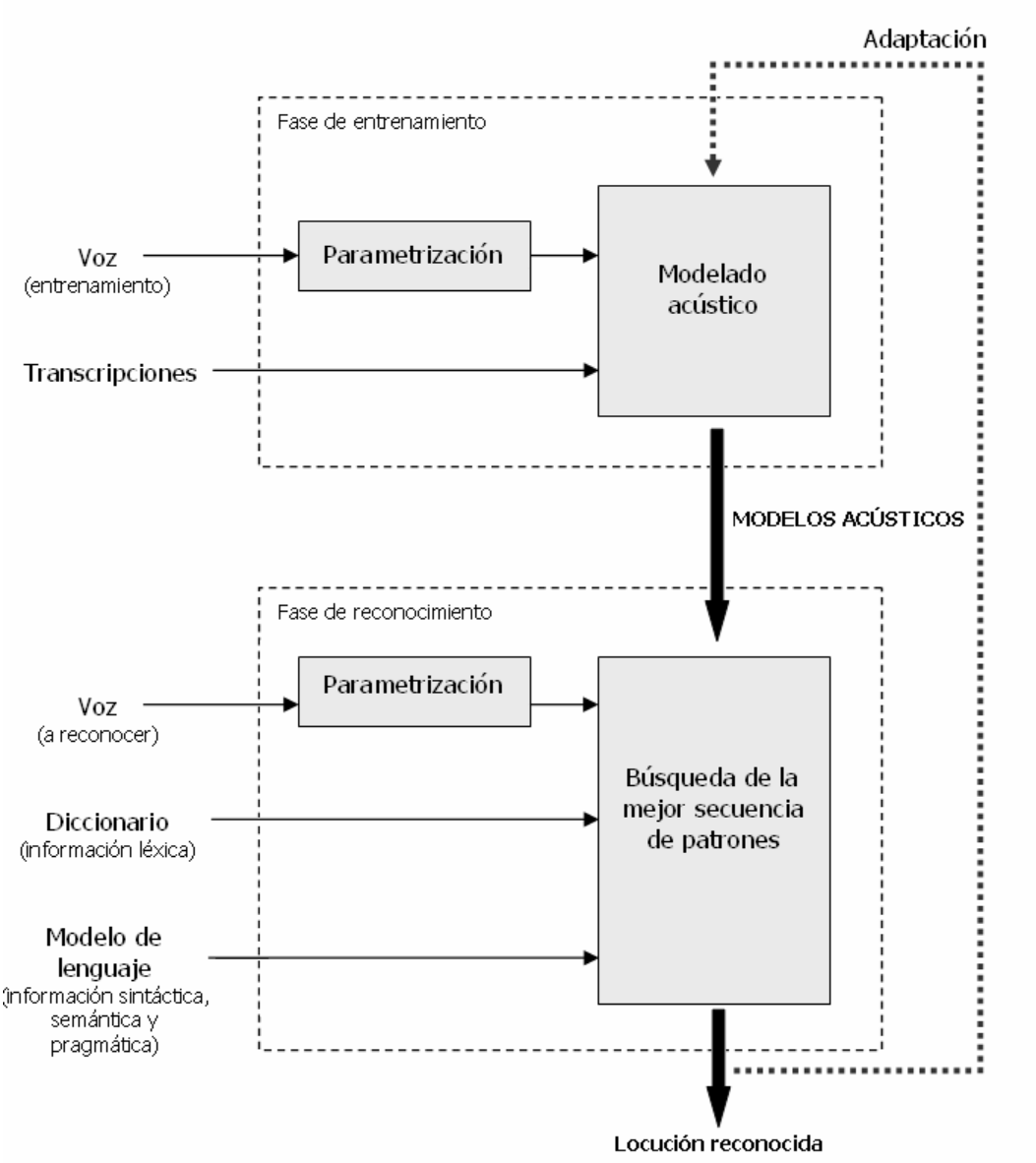

Figura 7.- Etapas de un sistema de reconocimiento automático de habla

Como se puede ver en la Figura 7, la primera fase será la de entrenamiento, que se encarga de obtener el modelado de los patrones de referencia, para lo que son necesarias unas locuciones de entrenamiento y sus transcripciones. Será en la fase de reconocimiento donde se busque la sucesión de patrones más adecuada para la señal de voz de entrada, utilizando para ello los patrones entrenados en la fase anterior, junto con información (ya sea léxica, sintáctica, semántica y/o pragmática, en función del reconocedor) que permita facilitar la búsqueda de la mejor secuencia. Fuera de estas dos fases principales, la salida del reconocedor puede utilizarse para realizar una adaptación de los modelos acústicos, cuyo objetivo es la búsqueda de una mejor representación de ciertas características, como el género del locutor, el locutor o el entorno. Esta adaptación también se puede realizar con nuevos datos de entrenamiento que no provengan de un reconocimiento previo.

En los siguientes apartados se pretende estudiar más detenidamente cada una de las etapas mencionadas anteriormente.

2.4.1 Parametrización

Las características físicas de cada locutor provocan que las propiedades de los sonidos que produce al hablar sean únicas y dependientes de esas características. Además, si un locutor repite variar veces una misma palabra, es muy probable que las características de los sonidos que la componen sean diferentes en cada realización. Por ambos motivos, el reconocimiento automático debe basarse en la información común que presentan los sonidos y que permiten diferenciarlos a nivel perceptual.

Por todo lo anterior parece fundamental realizar un tratamiento sobre la señal de voz de entrada del reconocedor automático, de forma que se obtengan las características que permiten diferenciar a los distintos sonidos. Este procesado de la señal también permite reducir las necesidades de almacenamiento para estos sistemas, ya que disminuye el tamaño respecto de lo que ocupa la voz de entrada.

Este procesado de la señal de entrada es lo que se conoce como parametrización.

2.4.1.1 Modelado fuente-filtro

En la Figura 4 se tiene un diagrama de bloques del proceso de generación de habla, representado como una señal de excitación que atraviesa un filtro, obteniéndose como salida la modulación de la señal de entrada con la respuesta del filtro [HAH01]. La señal de excitación será un tren de pulsos en el caso de sonidos sonoros, o ruido en el caso de sonidos sordos. El filtro representa la respuesta de las distintas cavidades resonantes que intervienen en la generación del habla.

Según el diagrama representado en esta figura, la señal de voz $y(t)$ se puede obtener como la convolución entre la señal de excitación $x(t)$ y la respuesta impulsiva del filtro $h(t)$.

$$y(t) = x(t) * h(t) \qquad (2.1)$$

Es conocido que las distintas cavidades resonantes y los articuladores realizan un modelado de la frecuencia de la señal que tienen a su entrada, por lo que será interesante trabajar en el dominio de la frecuencia, para lo que habrá que realizar la transformada de Fourier de (2.1):

$$Y(\omega) = X(\omega) \cdot H(\omega) \qquad (2.2)$$

Por lo tanto la señal de voz se puede ver como el producto de la transformada de Fourier de la señal de excitación por la respuesta en frecuencia del filtro. Lo que interesa en un sistema de reconocimiento automático de habla es obtener las características frecuenciales del filtro, que son las que van a marcar la diferencia entre los distintos sonidos a reconocer.

2.4.1.2 Enventanado de la señal de voz

El proceso descrito en el apartado anterior se basa en que puede realizarse la transformada de Fourier de (2.1). Esto es cierto para sistema lineales invariantes en el tiempo, donde las características permanecen estables a lo largo del tiempo, lo que no sucede en el caso de generación de voz, siendo éste un sistema lineal variante en el tiempo. Para poder asumir (2.2) es necesario separar la señal de voz en diferentes segmentos, dentro de los cuales se pueda considerar como estacionaria [HAH01], [YEG+06]. Por lo tanto, el enventanado es un proceso previo y necesario para la parametrización.

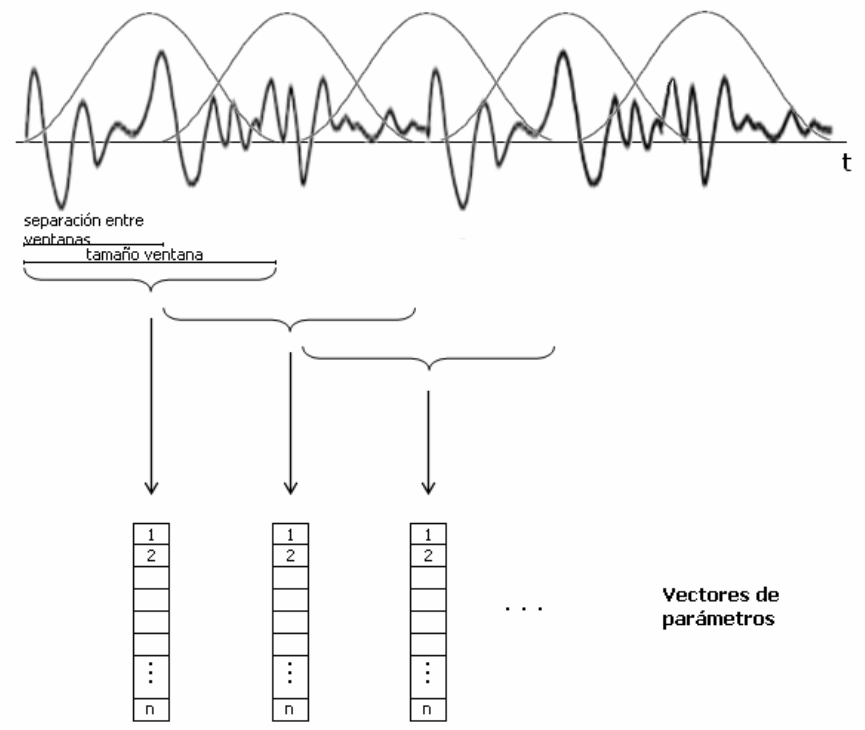

Figura 8.- Enventanado necesario para la parametrización

El tamaño de la ventana debe elegirse de tal forma que se cumpla la condición de estacionariedad. El límite superior estará marcado por la velocidad con la que el tracto vocal puede cambiar su forma, debiendo elegir un tamaño de ventana lo suficientemente pequeño como para considerar que la forma del tracto vocal no cambia. El límite inferior estará marcado por la resolución mínima aceptable en el dominio de la frecuencia, ya que a menor tamaño de ventana peor resolución frecuencial. En los sistemas de reconocimiento automático de habla se suele tomar un valor en el rango de 20 a 30ms.

Otra elección a realizar es la forma de la ventana. Se podría elegir una ventana rectangular, que implica simplemente separar los segmentos de la señal de voz, pero parece que la elección de otros tipos, como la ventana Hamming (2.3), atenúan la discontinuidad en los límites de los segmentos, lo que mejora los resultados obtenidos.

$$y_n^{'} = \left\{0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} y_n \qquad (2.3)$$

Donde $\{y_n, n = 1, N\}$ son las muestras de la señal, siendo N la longitud de la ventana.

Por último, mencionar que se considera beneficioso realizar un solape entre los distintos segmentos, para no perder características presentes en los bordes.

En la Figura 8 quedan representados todos los conceptos tratados en este apartado.

2.4.1.3 Procesamiento previo

Aunque no formen parte de un tipo de parametrización en concreto, existen algunas técnicas que, aplicadas con anterioridad, pueden mejorar las características de los parámetros obtenidos en cuanto a resultados de reconocimiento se refiere.

Entre estas técnicas se tienen:

- Eliminación de la componente media de continua de la señal de voz [YEG+06]: la voz con la que se trabaja es una señal digital, que proviene de una conversión analógico/digital. Este proceso de conversión puede introducir un offset de continua que no produce ningún beneficio para el reconocimiento, por lo que parece adecuado eliminarlo de la señal de voz. Este proceso se realiza para cada segmento en los que se divide la señal de voz para la parametrización.

- Filtro de preénfasis [HAH01]: este tipo de filtros aumenta la amplitud de las altas frecuencias, que se ven más atenuadas en el proceso de generación de voz que las bajas frecuencias. En reconocimiento de voz es habitual aplicar una ecuación diferencial de primer orden (2.4), sobre las muestras de cada segmento de voz.

$$y'_n = y_n - k \cdot y_{n-1} \qquad (2.4)$$

Siendo k el coeficiente de preénfasis, con un valor comprendido entre 0 y 1.

Todas estas técnicas se aplican sobre cada segmento de voz por separado, y antes de aplicar la ventana que se vaya a utilizar.

2.4.1.4 Cepstrum

El procesado cepstral [HAH01] permite separar, con cierta facilidad, la parte de la señal de voz que pertenece a la excitación y la parte que pertenece a la envolvente (respuesta del filtro).

El cepstrum de una señal [FUR89] se calcula haciendo la transformada inversa de Fourier (o similar) del logaritmo de la amplitud del espectro.

Para la señal de voz (ver (2.2)), el logaritmo de la amplitud se calcularía según (2.5):

$$\log|Y(\omega)| = \log|X(\omega)| + \log|H(\omega)| \qquad (2.5)$$

Y calculando la transformada inversa de Fourier se tiene (2.6):

$$c(\tau) = F^{-1}\log|Y(\omega)| = F^{-1}\log|X(\omega)| + F^{-1}\log|H(\omega)| \qquad (2.6)$$

Siendo $c(\tau)$ los coeficientes cepstrales, conocidos como *quefrencies*, y midiéndose en unidades temporales.

Con el proceso anterior se ha conseguido llegar a una suma de dos términos, siendo el primero de ellos proporcional a la excitación (estructura fina del espectro) y el segundo a la respuesta del filtro (envolvente espectral). En el dominio cepstral, el primer término se corresponde con una concentración a altas quefrencies, y el segundo término con una concentración a bajas quefrencies.

Para separar la información que caracteriza a la envolvente espectral de la que caracteriza a la fuente (proceso conocido como liftering) simplemente se deberá buscar el umbral de separación adecuado.

2.4.1.5 Análisis de banco de filtros

El sistema auditivo humano se comporta como un analizador espectral, es decir, un banco de filtros, con una respuesta no lineal para distintas frecuencias. Éstas son las características del oído que se van a simular en este tipo de parametrización, obteniéndose buenos resultados para los sistemas de reconocimiento de habla.

Partiendo de la señal de voz enventanada, se realiza la transformada de Fourier para tener su representación espectral. Esta señal se pasa a través del banco de filtros, de forma que para cada uno de los filtros se multiplica cada coeficiente de la señal por la amplitud del filtro correspondiente a esa frecuencia, y se realiza un sumatorio del resultado de esas multiplicaciones, obteniéndose así el coeficiente que representará a ese filtro en el vector de parámetros. Por cada ventana de voz se obtendrá un vector de parámetros de una longitud igual al número de filtros que formen el banco de filtros.

Es fundamental elegir bien este banco de filtros para el objetivo propuesto. Normalmente se trata de filtros triangulares solapados, con la caída a 3dB coincidiendo con la de los filtros adyacentes.

Existen varias opciones en cuanto a la escala utilizada [FUR89], [HAH01]:

- Escala lineal: las frecuencias centrales de cada filtro están equiespaciadas en frecuencia. Se le da el mismo valor a las frecuencias bajas que a las altas frecuencias.
- Escala logarítmica: se trabaja sobre el logaritmo de la frecuencia, que se adapta mejor al comportamiento del oído. Las frecuencias centrales de los filtros estarán equiespaciadas, pero esta vez en escala logarítmica, dándosele más peso a las bajas frecuencias que a las altas.
- Escala Bark y escala Mel: son escalas que surgen del concepto de banda crítica. La cóclea se puede ver como un banco de filtros solapados con anchos de banda igual a estas bandas críticas, a las que intentan adaptarse las escalas Bark y Mel. En ambas escalas la percepción es más fina a bajas frecuencias. La más utilizada en reconocimiento de habla es la escala Mel, que puede verse como una escala lineal hasta la frecuencia de 1KHz y logarítmica a partir de ese valor. En (2.7) se muestra cómo obtener la frecuencia mel a partir de la frecuencia lineal.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.7)$$

En la Figura 9 se representa la forma de un banco de filtros triangulares solapados y equiespaciados en escala mel. Gráficamente se puede observar que el ancho de banda de cada filtro es diferente según la frecuencia lineal, pero si se realiza la transformación a escala mel, en ella todos los filtros tendrán el mismo ancho de banda.

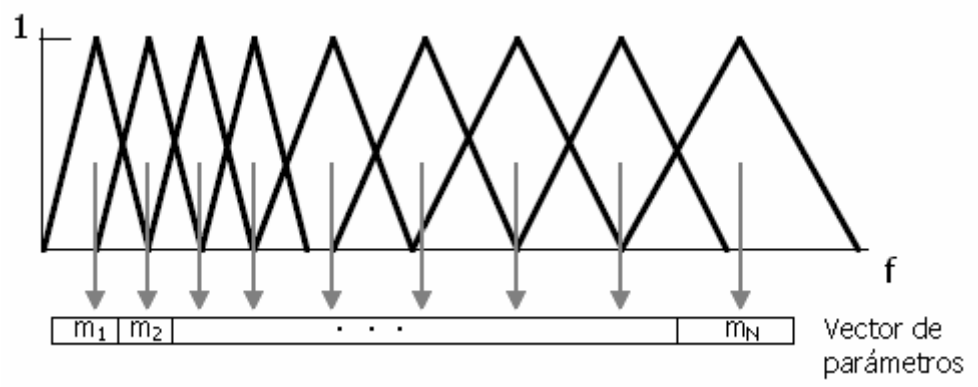

Figura 9.-Banco de filtros equiespaciados en escala Mel y el vector de parámetros correspondiente, siendo N el número de filtros

a) MFCCs ("Mel-Frequencies Cepstrum Coefficients")

Los parámetros obtenidos con la técnica anterior están muy correlados. Para evitar esto es recomendable el uso de una transformación al dominio cepstral de los parámetros del banco de filtro, lo que se conoce como coeficientes MFCCs o mel-cepstrum [YEG+06].

Partiendo de las amplitudes del banco de filtro (m_i) se calcula el cepstrum: logaritmo de dichas amplitudes y transformación en frecuencia (en este caso transformada discreta de coseno), proceso que queda resumido en (2.8):

$$c_i = \sqrt{\frac{2}{N}\sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right)} \qquad (2.8)$$

Donde $\{c_i, i = 1, M\}$ son los coeficientes cepstrales de un total de M coeficientes, y $\{m_j, j = 1, N\}$ son las amplitudes para el banco de N filtros. El valor de M debe ser suficientemente alto para que los coeficientes cepstrales contengan toda la información del tracto vocal, y lo suficientemente pequeño para no incluir la información de la señal de excitación.

Este tipo de parametrización es el más usado en los sistemas de reconocimiento de habla, y en particular en el sistema analizado en el presente proyecto fin de carrera.

b) CMN ("Cepstral Mean Normalization")

Si se tiene en cuenta que la señal de voz que llega al reconocedor de habla no es la misma que pronuncia el locutor, sino que se ha visto modificada por el canal de transmisión, parece deseable eliminar esta información puesto que no va a aportar nada favorable al reconocimiento [YEG+06]. Los coeficientes MFCCs permiten realizar este proceso de una manera muy simple: se trata de calcular, para cada coeficiente por separado, su media a lo largo de todos los vectores de parámetros, y eliminarla del valor inicial. Este proceso se repetirá para cada locución de entrenamiento y test por separado, conociéndose como técnica CMN.

2.4.1.6 Análisis lineal predictivo

El aparato fonador humano se puede ver como una concatenación de tubos sin pérdidas, a través de los cuales viaja la señal de excitación[HAH01]. Esta característica es la que se utiliza en este tipo de parametrización para representar a la señal de voz.

Experimentalmente se ha obtenido que un filtro con una función de transferencia todo polos en el dominio de la transformada z (2.9) modela bastante bien el sistema anterior, estando relacionados los coeficientes del filtro con los coeficientes de reflexión que se producirían en la unión entre los distintos tubos, y que se calculan de forma que se minimice el error de predicción total. El vector de parametrización obtenido tendrá una dimensión igual al número de tubos con los que se modele el tracto vocal.

$$H(z) = \frac{1}{1 - \sum_{j=1}^{N} a_j z^{-j}} \qquad (2.9)$$

Donde $\{a_j, j = 1, N\}$ son los coeficientes del filtro (LPC), con N representando el número de tubos.

Igual que en el caso de análisis de banco de filtros, aquí también es habitual transformar estos parámetros al dominio cepstral.

2.4.1.7 Técnicas a posteriori

Independientemente de la parametrización utilizada se puede añadir más información al vector de parametrización para intentar mejorar los resultados de reconocimiento [YEG+06].

Una de las técnicas consistiría en añadir información sobre la energía de la señal, ya sea el coeficiente de autocorrelación en la posición 0 para parámetros LPC, o bien el logaritmo de la energía de la señal, normalizada o no y calculada antes o después del enventanado y del preénfasis.

Otra técnica muy utilizada consiste en utilizar derivadas temporales de los parámetros estáticos iniciales: de primer orden o coeficientes delta, de segundo orden o coeficientes de aceleración y coeficientes de regresión de tercer orden.

En (2.10) se muestra la forma de obtener los coeficientes delta a partir de los coeficientes estáticos, que será la misma que se aplique para obtener las aceleraciones a partir de los coeficientes delta, o para obtener los coeficientes de tercer orden a partir de las aceleraciones.

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (2.10)$$

Donde d_t es el coeficiente delta correspondiente al coeficiente estático c_t.

2.4.2 Modelado acústico

El modelado acústico es la parte más importante del reconocedor de habla y se debe elegir de forma que cumpla con la precisión de reconocimiento esperada.

El modelado acústico a desarrollar en un sistema de reconocimiento de voz estará condicionado por la técnica de reconocimiento de habla que se vaya a utilizar. Después de explicar las distintas técnicas existentes, los siguientes subapartados se centrarán en explicar el modelado acústico para la técnica de reconocimiento elegida para el desarrollo del presente proyecto fin de carrera.

2.4.2.1 Técnicas de reconocimiento

Como ya se ha comentado con anterioridad, en un sistema de reconocimiento se intenta obtener la secuencia de símbolos oculta bajo una señal de voz de entrada. Esta señal de voz es transformada en una secuencia de vectores de parámetros equiespaciados, considerándose que cada vector de parámetros cubre las características más relevantes de un segmento de voz que se considera estacionario.

El trabajo del reconocedor será realizar un mapeo entre esta secuencia de vectores y la secuencia de símbolos inicial. Por lo tanto, la tarea de reconocimiento puede verse como una tarea de clasificación de patrones, siendo el mensaje reconocido una concatenación de patrones.

Las principales técnicas de clasificación de patrones son las siguientes [Col01], [Lli09]:

- DTW (*Dynamic Time Warping*): esta técnica aplica programación dinámica a la comparación de patrones. Se parte de una serie de plantillas entrenadas previamente que identifican las características de los patrones a reconocer. La voz a reconocer se comparará con estas plantillas, realizando un alineamiento no lineal en el tiempo para paliar los efectos de expansiones o compresiones temporales de los patrones, y se elegirá como salida aquel patrón que ofrezca una menor distancia con la señal de entrada.
- Redes Neuronales: estas redes pretenden simular la flexibilidad computacional del cerebro humano, lo que puede ser fácilmente adaptado al reconocimiento de habla.
  El cerebro está dotado de una gran malla de neuronas interconectadas, con billones de conexiones que permiten que se comuniquen intercambiando distintos impulsos nerviosos. Estos impulsos nerviosos, sin sentido cuando se les considera de forma aislada, toman su significado cuando son considerados como un conjunto con el resto de impulsos. Por similitud, podemos decir que estos impulsos nerviosos son para el cerebro lo que los fonemas para el lenguaje: no tienen sentido a no ser que se los considere con el resto de fonemas.
  Existen técnicas de reconocimiento de habla aplicando redes neuronales, que gracias a la topología que utilizan pueden aplicar técnicas muy simples operando en paralelo, buscando conseguir resultados en tiempo real.
- HMMs (Hidden Markov Models): Es una forma de modelado estadístico, tratando de modelar los distintos patrones como procesos estocásticos. Se parte de distintas observaciones de cada modelo, que deberán caracterizar todas las posibles variaciones del mismo, y para cada uno se construye un modelo paramétrico con el que se hará la clasificación de patrones ayudados de un algoritmo de alineación temporal no lineal (programación dinámica).

Como puede deducirse los principios son los mismos que para la DTW, pero difiere en la forma de obtener los patrones, el tipo de patrón, las medidas que se realizan para decidir el patrón adecuado e incluso en el alineamiento temporal.

Se han convertido en uno de los métodos estadísticos más potentes para el modelado de señales de voz.

- Segment Models: se trata de modelos que pretenden solucionar dos problemas presentes en los HMMs:
  - Para obtener unos HMMs que modelen correctamente la señal de voz es necesario partir de segmentos de voz con características estacionarias, que en ocasiones no se consigue.
  - Al elegir la estructura de un HMM se dejan ciertos grados de libertad que no son modelados en el entrenamiento, y que en el reconocimiento pueden provocar la elección de modelos equivocados.

  Las características de estos nuevos modelos son semejantes a los HMMs, pero modelando el tiempo de una manera más elaborada.

El reconocimiento de voz basado en DTW es fácil de implementar y rápido para reconocimiento de voz de pequeño vocabulario. Gracias al alineamiento temporal obtenido por la programación dinámica se puede acotar diferencias entre locutores o incluso entre repeticiones de un mismo patrón para un determinado locutor. Sin embargo, no se puede obtener un único modelo global para representar todas las características que puede tener un mismo modelo en distintas instancias del mismo, sino que se necesitan distintos modelos para caracterizar todas estas diferencias. Los HMMs se presentan como una mejor alternativa bajo este punto de vista.

En cuanto a las redes neuronales, su desventaja principal es la gran carga computacional que conllevan, por lo que no han sido utilizados ampliamente hasta ahora. Además, necesitan de un entrenamiento muy costoso, puesto que no se conoce lo estructura interna ni el número de nodos de red necesarios, y tienen la posibilidad de no obtener la solución óptima.

Parece ser que las redes neuronales presentan soluciones muy efectivas si se combinan con otras técnicas (HMMs o DTW), que palian la incapacidad de las redes neuronales para el procesado de la información temporal de los patrones, ya que estas redes por si solas únicamente realizan un procesado espacial. Existen otras soluciones que añaden este tratamiento en las propias redes neuronales (incorporación de memoria) pero aumenta mucho su dificultad.

En cuanto a los Segment Models se han obtenido muy buenos resultados en tareas de reconocimiento de vocabulario pequeño o de palabras aisladas, pero no se usa en reconocimiento de habla continua con un vocabulario grande debido a la gran complejidad de implementación de estos modelos.

Por todo lo anterior, aunque parece que en reconocimiento de palabras aisladas o en reconocedores con un vocabulario pequeño todas las técnicas pueden dar buenos resultados, de la única de la que se tienen evidencias de su buen comportamiento en reconocimiento de habla continua con un vocabulario extenso es del uso de HMMs. En el presente proyecto fin de carrera se pretende trabajar con reconocimiento de habla continua, por lo que a partir de este momento sólo se va a estudiar con profundidad la técnica de los HMMs.

2.4.2.2 Reconocimiento de palabras aisladas

Para explicar las características del modelado acústico utilizando HMMs se va a partir del caso más sencillo posible para después extenderlo al que realmente se usará en este proyecto fin de carrera. Por lo tanto, se partirá del reconocimiento de palabras aisladas, contando con un HMM que caracterice a cada palabra, y se llegará a un reconocimiento de habla continua con HMMs referidos a fonemas.

En el caso de reconocimiento de palabras aisladas, se va a partir de la regla fundamental de reconocimiento (teoría de decisión de Bayes) para introducir lo que se debe caracterizar al entrenar un HMM, y así poder utilizarlo con posterioridad en el reconocimiento. Junto con esto se explicará la topología de un HMM y los algoritmos utilizados para entrenarlos y para realizar el reconocimiento.

a) Teoría de decisión de Bayes

La teoría de decisión de Bayes [Lli09], [YEG+06] es la base del reconocimiento de patrones estadísticos. Puesto que los HMMs son patrones estadísticos es necesario comenzar por conocer las bases de esta teoría.

El problema de reconocimiento estadístico se puede resumir en encontrar la palabra perteneciente al vocabulario del reconocedor que ofrezca una mayor probabilidad dada la observación de entrada. Este principio fundamental de reconocimiento se puede expresar como sigue (2.11):

$$\arg\max_i \{P(\omega_i|O)\} \qquad (2.11)$$

Siendo ω_i la i-ésima palabra del vocabulario del reconocedor, y $O = o_1, o_2, ..., o_T$ la observación de entrada, con o_t el vector de voz parametrizado observado en el instante t y T el número de vectores de parámetros que componen esa observación de entrada.

La probabilidad anterior no se puede calcular directamente. Aplicando la regla de Bayes a esta probabilidad se tiene que (2.12):

$$P(\omega_i|O) = \frac{P(O|\omega_i) \cdot P(\omega_i)}{P(O)} \qquad (2.12)$$

Siendo $P(O) = \sum_i P(O|\omega_i) \cdot P(\omega_i)$.

Si se quiere buscar la palabra ω_i que maximice (2.12), hay que tener en cuenta que el término del denominador se mantiene constante. Si además se utiliza un conjunto de palabras equiprobables, la palabra ω_i buscada será la que cumpla (2.13):

$$\arg\max_i \{P(O|\omega_i)\} \qquad (2.13)$$

La fórmula (2.13) se conoce como teoría de decisión de Bayes.

El cálculo directo de la probabilidad condicional anterior no es manejable. Sin embargo, si se dispone de un modelo paramétrico, M_i que represente a dicha palabra, ω_i, el problema de reconocimiento se reduce al problema mucho más simple de cálculo de $P(O|M_i)$, función de los parámetros de dicho modelo.

b) Definición de HMM

Un modelo de Markov [YEG+06], [Lli09] es una máquina de estados finitos, que en momentos temporales equiespaciados cambia de un estado i a otro j en función de una probabilidad de transición a_{ij}, y que al llegar en un momento t a un estado emisor j emitirá un vector de observación o_t con una densidad de probabilidad $b_j(o_t)$.

Para ilustrar los conceptos anteriores, en la Figura 10 se muestra un modelo de Markov de 6 estados, con el primero y el último no emisores (característica necesaria para reconocimiento de habla continua), donde se representan distintas transiciones entre estados y densidades de probabilidad de emisión.

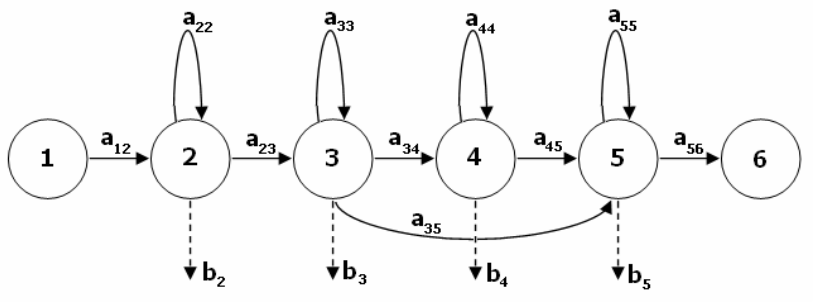

Figura 10.- Ejemplo de modelo de Markov

Para explicar el principio de funcionamiento de estos modelos, se va a suponer un ejemplo donde la secuencia de observaciones $O = o_1, o_2, o_3, o_4, o_5, o_6$ es generada por el modelo anterior con una secuencia de estados $O = 1,2,2,3,4,4,5,6$. Con estos datos, la probabilidad conjunta de que la secuencia de observaciones O sea generada por el modelo M dada la secuencia de estados X será (2.14):

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)a_{34}b_4(o_4)a_{44}b_4(o_5)a_{45}b_5(o_6)a_{56} \qquad (2.14)$$

En el caso de reconocimiento de voz, la secuencia de estados que siguen las observaciones no es conocida, motivo por el que los modelos se conocen como modelos ocultos de Markov (HMM). Por lo tanto, para obtener la probabilidad de que la observación O sea generada por el modelo M se tendrá que calcular la probabilidad anterior, (2.14), para cada secuencia de estados posibles que puedan generar esa secuencia de observaciones y realizar el sumatorio de todas ellas, según se muestra en (2.15).

$$P(O|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \qquad (2.15)$$

Suponiendo que el primero y el último de los estados de la secuencia ($x(0), x(T+1)$) son no emisores, y que la dimensión de la secuencia de observaciones es T.

Una aproximación a (2.15) es asignarle el valor que se obtiene con la secuencia de estados que proporcione el valor máximo a la probabilidad condicional conjunta, como se muestra en (2.16).

$$\hat{P}(O|M) = \max_X \left(P(O, X|M)\right) = \max_X \left(a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right) \qquad (2.16)$$

Existen procedimientos más o menos simples para obtener el valor de $P(O|M)$ que se estudiarán en los apartados dedicados a ello. Por ahora simplemente es necesario notar que si se puede caracterizar la probabilidad de una secuencia de observaciones para un determinado modelo, se tiene resuelto el problema de reconocimiento, y como se ha visto, cada modelo quedará caracterizado por los valores de sus transiciones, $\{a_{ij}\}$, y de las densidades de probabilidad de emisión de cada estado, $\{b_j(o_t)\}$, que se deberán obtener en una fase previa de entrenamiento.

Como se puede observar en la Figura 10, existen distintas variables a fijar en la topología de un HMM:

- Número de estados: dependerá del uso que se le quiera dar al modelo. Si cada modelo va a representar a un fonema, son necesarios de 3 a 5 estados emisores. Sin embargo, si cada modelo va a representar a una palabra son necesarios más estados, debido a que las muestras de voz que van a representar tienen una duración mayor (hay que permitir más variabilidad que en el caso de fonemas).
- Transiciones entre estados: hay que elegir las transiciones permitidas entre los distintos estados. También será función de lo que se quiera representar en cada modelo. Por ejemplo, en el caso de los silencios se pueden elegir transiciones que permitan saltarse estados, para modelar así silencios más cortos.
- Forma de las densidades de probabilidad de emisión de cada estado: se pueden elegir funciones discretas, para el caso de que los datos procedan de un conjunto de datos finitos, o funciones continuas para cuando proceden de un espacio continuo, tratándose de HMMs discretos en el primer caso y HMMs continuos en el segundo. En reconocimiento de habla normalmente se usan HMMs continuos, modelándose las densidades de probabilidad de emisión como mezclas de gausianas multidimensionales. Para este caso concreto, una vez elegida la forma que va a tomar la función, habrá que elegir la dimensión de las gausianas, el número de gausianas en la mezcla y los pesos que se le asignan a cada una.

  Existe un tercer tipo de HMMs, los HMMs semicontinuos, mezcla de los dos anteriores, para los que se tienen un conjunto finito de densidades de probabilidad continuas que serán asignadas a cada HMM con un cierto peso.

  Cuando se tienen suficientes datos de entrenamiento, los HMMs continuos ofrecen mejores resultados que los HMMs discretos o semicontinuos. Cuando los datos son insuficientes o bien el número de mezclas es bajo, los HMMs continuos empiezan a trabajar peor que los otros dos tipos.

En el uso de HMMs en reconocimiento de habla es habitual elegir una topología de izquierda a derecha (como la mostrada en la Figura 10), donde una vez alcanzado un estado no se permite volver a los estados anteriores. Se elige esta topología porque se corresponde

bastante bien con el carácter no estacionario de la señal de voz que se intenta modelar con cada HMM. Cada estado representa que se alcanza un segmento cuasi-estacionario y para modelar características similares de la voz contigua a este segmento se añade la transición a sí mismo. Para permitir que este segmento evolucione, se añade transición al estado siguiente. Transiciones a estados anteriores representarían una evolución no natural de la señal de voz. Si estas últimas no se añaden, el movimiento dentro del modelo se realiza de izquierda a derecha, lo que le da el nombre a esta topología, la más usada en el estado del arte del reconocimiento de habla.

c) Simplificación de los HMMs

Cuando se usan HMMs, normalmente se hacen dos simplificaciones que facilitan el entrenamiento y la evaluación de estos modelos [Lli09], puesto que reducen en gran medida el número de parámetros a calcular, sin pérdida significativa de la capacidad de modelado.

La primera simplificación es conocida como *asunción de Markov* e implica que la probabilidad de estar en un estado en un momento t sólo depende del estado en el que se encontraba en el momento anterior, y no de la secuencia de estados seguida hasta llegar a él.

La segunda simplificación es conocida como *asunción de independencia de salida* e implica que la probabilidad de emisión de una determinada observación en el momento t sólo depende del estado en el que se encuentra en ese momento, siendo independiente de la secuencia de estados recorridos y de las observaciones emitidas en ese recorrido.

Los modelos que presentan estas simplificaciones se presentan como *HMMs de primer orden*, y son los que se utilizarán a partir de ahora.

d) Estimación de parámetros

Si se tiene un conjunto de ejemplos de una determinada palabra, se puede generar un HMM que modelará a esa voz real. Cuanto más representativo sea el conjunto de muestras, más fuentes de variabilidad quedarán representadas en dicho modelo.

En este subapartado se va a explicar el procedimiento usado para obtener los parámetros de un HMM [YEG+06], partiendo de que las densidades de probabilidad de salida serán mezclas de gausianas multidimensionales.

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} \mathrm{N}(o_t; \mu_{jm}; \Sigma_{jm}) \qquad (2.17)$$

Donde $b_j(o_t)$ es la función de densidad de probabilidad de emitir la observación en el instante t, o_t, por el estado j, siendo c_{jm} el peso de la m-ésima gausiana multidimensional de la mezcla de M gausianas que ocupan el estado j, y $\mathrm{N}(o;\mu_{jm};\Sigma_{jm})$ la función de cada gausiana de la mezcla de ese estado (2.18), con un vector media μ_{jm} y una matriz de covarianza Σ_{jm}.

$$\mathrm{N}(o;\mu;\Sigma)=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \qquad (2.18)$$

Como se muestra en la Figura 11, un estado formado por una mezcla de gausianas no es más que un conjunto de subestados de una única gausiana. Por sencillez, y gracias a esta característica, se va a explicar la estimación de los parámetros para el caso simple de estados con una única gausiana, puesto que es generalizable a mezclas de M gausianas.

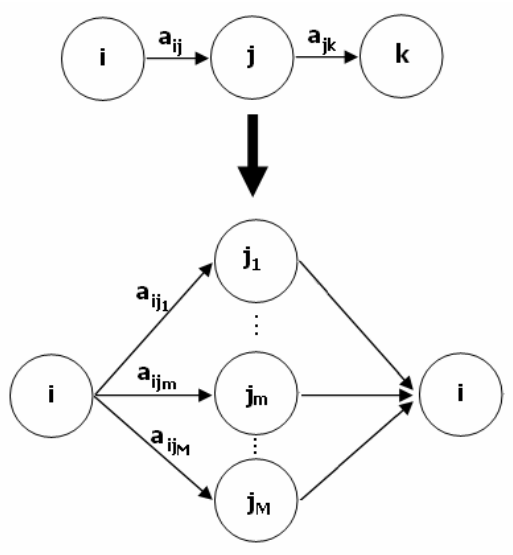

Figura 11.- Mezcla de M gausianas

Por lo tanto, a partir de este momento, cada estado j estará modelado por una única gausiana.

Una vez elegida la forma de las funciones de densidad de probabilidad de emisión, debe darse un valor inicial a sus parámetros. Las fórmulas generales de cálculo de estos parámetros son las siguientes (2.19), (2.20):

$$\hat{\mu}_j = \frac{1}{N}\sum_{n=1}^{N} o_n \qquad (2.19)$$

$$\hat{\Sigma}_j = \frac{1}{N}\sum_{n=1}^{N} (o_n - \mu_j)(o_n - \mu_j)' \qquad (2.20)$$

Siendo N el número de vectores de observaciones emitidos por el estado j, y $\{o_n\}$ el conjunto de observaciones emitidos por dicho estado.

Como los vectores de observaciones emitidos por cada estado no son conocidos, ya que la secuencia de estados es oculta, es necesario realizar una aproximación de estos valores. Uno de los procedimientos más extendidos de inicialización (que será el utilizado en el presente proyecto fin de carrera) consiste en suponer que todos los vectores de observaciones pertenecen al mismo estado, calcular los valores a partir de (2.19) y (2.20), y asignar los resultados a las medias y covarianzas de todos los estados de todos los modelos.

Existen otros procedimientos usados en el caso de reconocimiento de palabras aisladas, aunque no se va a entrar en detalle debido a que se escapan del ámbito del proyecto.

Una vez obtenido un valor inicial para los parámetros, se procede a una reestimación de los mismos para ajustarse a los datos de entrenamiento, (2.21) y (2.22), utilizándose para ello el *algoritmo de Baum-Welch*.

Se parte de una muestra de entrenamiento, $O = o_1, o_2, ..., o_T$, para un determinado modelo, donde o_t indica el vector de observaciones para el instante t con $t = 1,...,T$, siendo T la posición del último vector de observación para dicha muestra.

Para calcular la probabilidad de esa observación dado el modelo M_i, como no se conoce la secuencia de estados, es necesario sumar las probabilidades de todas las secuencias de estados. Por lo tanto, se tendrá la contribución de cada o_t en cada uno de los estados, contribución que se verá escalada por la probabilidad de estar en ese estado al producirse dicha observación.

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t) o_t}{\sum_{t=1}^{T} L_j(t)} \qquad (2.21)$$

$$\hat{\Sigma}_j = \frac{\sum_{t=1}^{T} L_j(t)(o_t - \mu_j)(o_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)} \qquad (2.22)$$

Donde $L_j(t)$ indica la probabilidad de estar en el estado j en el instante t.

Por lo tanto, se necesita obtener $L_j(t)$, para lo que se utiliza el *algoritmo forward-backward*, que se basa en dos términos conocidos como probabilidad forward, $\alpha_j(t)$, y probabilidad backward, $\beta_j(t)$, que necesitan ser explicados previamente al algoritmo.

Dado un modelo M con un total de N estados, la probabilidad forward se define como la probabilidad de que se hayan dado los t primeros vectores de observación y estar en el estado j en ese instante t (2.23):

$$\alpha_j(t) = P(o_1, o_2, ..., o_t, x(t) = j | M) \qquad (2.23)$$

Esta probabilidad conjunta se puede calcular como la probabilidad de emitir la observación o_t dado que se está en el estado j por la probabilidad de estar en ese estado en el instante t, que se puede calcular como sumatorio de todas las probabilidades forward de estar en cualquier estado en el momento anterior por la probabilidad de transición de ese estado al actual. Por lo tanto, se tiene la recursividad (2.24):

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij} \right] b_j(o_t) \qquad (2.24)$$

Hay que notar que el sumatorio anterior sólo va del estado 2 al $N-1$, debido al caso particular de que en los modelos que se van a usar, los estados primero y último son no emisores.

Las condiciones iniciales se muestran en (2.25) y (2.26):

$$\alpha_1(1) = 1 \qquad (2.25)$$

$$\alpha_j(1) = a_{1j} b_j(o_1) \qquad (2.26)$$

Y la condición final se muestra en (2.27):

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN} \qquad (2.27)$$

Esta condición final, es decir, la probabilidad de estar en el estado final dado que se han emitido todas las observaciones, se corresponde con $P(O|M)$, la probabilidad de tener la observación O dado el modelo M.

En cuanto a la probabilidad backward, dado el modelo M anterior, se define como la probabilidad de que estando en el estado j en el instante t, se genere la secuencia de observaciones restantes (2.28):

$$\beta_j(t) = P(o_{t+1}, o_{t+2}, ..., o_T | x(t) = j, M) \qquad (2.28)$$

Esta probabilidad se puede calcular también recursivamente como sumatorio para todos los estados, de la probabilidad de emisión de la siguiente muestra por cualquiera de esos estados, teniendo en cuenta las probabilidades de transición existentes desde el estado inicial j, por la probabilidad backward de que se genere la secuencia de observaciones de $o_{t+1}, ..., o_T$ a partir de ese nuevo estado en el que se encuentra en el instante $t+1$. Por lo tanto, se tiene la recursividad (2.29):

$$\beta_j(t) = \sum_{i=2}^{N-1} a_{ji} b_i(o_{t+1}) \beta_i(t+1) \qquad (2.29)$$

Igual que en el caso anterior el sumatorio sólo va del estado 2 al $N-1$, debido al caso particular de que los estados primero y último del modelo son no emisores.

La condición inicial se muestra en (2.30):

$$\beta_j(T) = a_{jN} \qquad (2.30)$$

Y la condición final se muestra en (2.31):

$$\beta_1(1) = \sum_{i=2}^{N-1} a_{1i} b_i(o_1) \beta_i(1) \qquad (2.31)$$

A partir de las definiciones de ambas probabilidades ((2.23) y (2.28)) se puede obtener una expresión (2.32) para la probabilidad de estar en el estado j en el instante t, es decir, para $L_j(t)$:

$$L_j(t) = P(x(t) = j|O, M) = \frac{P(O, x(t) = j|M)}{P(O|M)} = \frac{\alpha_j(t)\beta_j(t)}{P(O|M)} \qquad (2.32)$$

Una vez conocida la relación entre $L_j(t)$ y las probabilidades forward y backward, ya se puede completar la explicación del *algoritmo de Baum-Welch* para la reestimación de parámetros. Partiendo de la inicialización de los parámetros:

- Se calculan las probabilidades forward y backward.
- Se calcula $L_j(t)$ para cada estado j y cada instante t de la secuencia de vectores de observaciones que componen O.
- Se calcula el nuevo valor de los parámetros a partir de (2.21) y (2.22).
- Se calcula $P(O|M)$, y se vuelve a empezar de nuevo siempre que se obtenga un valor mayor al valor anterior.

La explicación anterior se ha dado para el caso de tener una única observación O para el modelo. En los casos reales se tiene un conjunto de observaciones para cada modelo, por lo que hay que extender la reestimación anterior a este nuevo caso. La única diferencia es que los tres primeros pasos descritos anteriormente se realizan para cada observación, obteniendo el valor final de los parámetros a partir de los valores obtenidos para esos parámetros en cada observación. A partir de ahí, el criterio de parada será el mismo, que $P(O|M)$ no crezca respecto a la iteración anterior.

Para obtener más información sobre la reestimación de las probabilidades de transición, consultar la referencia [YEG+06].

e) Descodificación acústica

Una vez obtenidos los parámetros de cada HMM (que en este caso representa a cada una de las palabras con la que trabaja el reconocedor), el problema de reconocimiento se basa [YEG+06] en, dada una observación de test, O, obtener el modelo, M_i, que proporcione una mayor probabilidad de dicha observación (2.33).

$$\arg\max_i \{P(O|M_i)\} \qquad (2.33)$$

Como se ha visto en el apartado anterior, es posible obtener $P(O|M_i)$ a partir de la probabilidad forward ($\alpha_N(T)$ para cada modelo M_i). Calculando esta probabilidad para cada modelo se tendría la palabra reconocida a través del modelo que ofrezca un mayor valor de dicha probabilidad.

Aunque esto es cierto, debido a que el objetivo fundamental de este apartado es ofrecer un punto de partida para el reconocimiento de habla continua, se va a enfocar el reconocimiento en la secuencia de estados que ofrecen la máxima probabilidad para cada modelo del conjunto, eligiendo el modelo con el mayor valor de dicha probabilidad.

Existe un algoritmo, conocido como *algoritmo de Viterbi*, que permite calcular la probabilidad anterior de forma recursiva, cuyos pasos se explican a continuación.

Con $\phi_j(t)$ se representa el valor máximo para la probabilidad de haber observado los vectores de observación de o_1 a o_t estando en el estado j en el instante t. Esta probabilidad se puede calcular como la probabilidad de emitir la observación o_t dado que se está en el estado j multiplicada por un segundo término que se calcula como el producto de:

- El valor máximo para la probabilidad de haber observado la secuencia de observaciones de o_1 a o_{t-1} estando en el estado i en el instante $t-1$.
- La probabilidad de transición del estado i al j.

Siendo i el estado que produce mayor valor de este segundo término.

Por lo tanto, se tiene la recursividad (2.34):

$$\phi_j(t) = \max_i \{\phi_i(t-1)a_{ij}\} b_j(o_t) \qquad (2.34)$$

Las condiciones iniciales se muestran en (2.35) y (2.36):

$$\phi_1(1) = 1 \qquad (2.35)$$

$$\phi_j(1) = a_{1j} b_j(o_1) \qquad (2.36)$$

Y la condición final se muestra en (2.37):

$$\phi_N(T) = \max_i \{\phi_i(T)a_{iN}\} \qquad (2.37)$$

Esta condición final, el valor máximo de la probabilidad de estar en el estado final dado que se han emitido todas las observaciones, se corresponde con una estimación de $P(O|M)$, la probabilidad de tener la observación O dado el modelo M. Es una estimación puesto que el valor real de esta probabilidad no sólo tiene en cuenta la secuencia de estados que proporciona el valor máximo, sino todas las secuencias de estados posibles.

Debido a que los términos de la recursión mostrada en (2.34) pueden ser demasiado pequeños y producir problemas de cálculo, el *algoritmo de Viterbi* trabaja aplicando el logaritmo a dicha fórmula. Además, esta representación (2.38) admite una representación visual del algoritmo que permite entender más claramente su funcionamiento.

$$\varphi_j(t) = \log(\phi_j(t)) = \max_i \{\varphi_i(t-1) + \log(a_{ij})\} + \log(b_j(o_t)) \qquad (2.38)$$

En la Figura 12 quedan representados todos los elementos que intervienen en el cálculo de $\varphi_j(t)$. El eje vertical representa los estados del modelo que se está evaluando, el eje horizontal indica los vectores de observación correspondientes a la muestra O que se está evaluando, los puntos indican las probabilidades de emisión ($b_j(t)$) y las líneas indican las probabilidad de transición entre estados (a_{ij}).

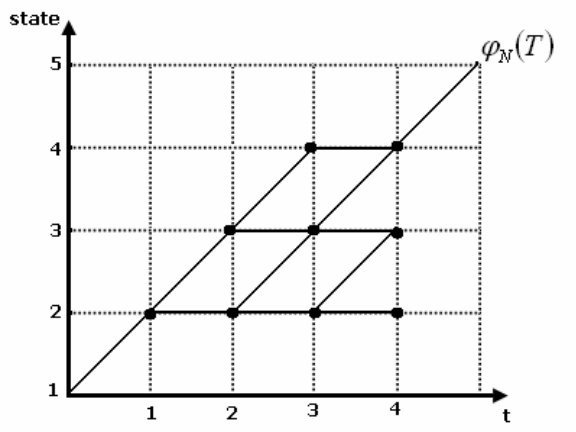

Figura 12.- Algoritmo de Viterbi

En el caso de la Figura 12 se tiene un HMM de 5 estados, con el primero y último estados no emisores, transiciones permitidas al mismo estado o al estado de la derecha y una muestra formada por 4 vectores de observación ($O = o_1, o_2, o_3, o_4$).

Si se quiere calcular $\varphi_j(t)$, se parte de la columna correspondiente a esa muestra de tiempo t y se evalúan todos los posibles caminos desde cualquier estado i (filas), sabiendo que todos los caminos crecen de izquierda a derecha. Se calcularía $\varphi_i(t-1)$ para todos los estados

i posibles, eligiéndose aquel que produzca mayor valor del producto $\varphi_i(t-1)a_{ij}$, multiplicándose después por $b_j(t)$.

2.4.2.3 Reconocimiento de habla continua

El reconocimiento de habla continua se diferencia del anterior en que tanto las locuciones de entrenamiento como las locuciones a reconocer no van a estar formadas por un único HMM (como era el caso de reconocimiento de palabras aisladas cuando los HMMs modelan a cada palabra del vocabulario) sino que estarán formadas por un conjunto concatenado de HMMs.

Los principios de entrenamiento de reconocedores de palabras aisladas se pueden aplicar en este caso, pero teniendo en cuenta que las locuciones de entrada se corresponden con varios modelos concatenados. Si se dispone de la información necesaria para poder aislar la parte de la locución que se corresponde con cada modelo, se podría entrenar cada modelo de forma aislada de la manera ya descrita.

En general esta información no está disponible, y si lo está, la cantidad de datos suele ser insuficiente. Debido a esta limitación, la aplicación del algoritmo de Baum-Welch necesita de algunas modificaciones para adaptarse a este entrenamiento:

- Todos los modelos se entrenan en paralelo.
- Para cada locución de entrenamiento (O):
  - Se crea un HMM compuesto, formado por la concatenación de los distintos HMMs que indique la trascripción de dicha locución. Para esto es indispensable que los estados de entrada y salida de cada HMM sean no emisores.
  - Se aplica el *algoritmo de Baum-Welch* a ese HMM compuesto:
    - Cálculo de las probabilidades forward y backward.
    - Cálculo de $L_j(t)$ para cada estado j y cada instante t de la secuencia de vectores de observaciones que componen O.
    - Se obtienen los parámetros de cada estado a partir de los valores anteriores.
- Una vez realizado el procedimiento anterior para cada locución, se actualizan los parámetros de cada estado de cada HMM a partir de los valores obtenidos para el mismo en las locuciones de entrenamiento en las que participe.
- Este procedimiento deberá ser repetido hasta conseguir la convergencia.

Para el reconocimiento se suele utilizar una simplificación del *algoritmo de Viterbi*. Esta simplificación se conoce como *Token passing model*.

Para el caso particular de reconocimiento de palabras aisladas, cada estado j, del modelo que se está evaluando, en un determinado momento t almacenará un token con diferentes informaciones, y en particular el valor de $\varphi_j(t)$.

El algoritmo *Token passing model*, en un momento t pasa el token almacenado en cada estado i para el momento anterior $t-1$ a todos los estados j conectados, modificando su valor en $\log(a_{ij}) + \log(b_j(o_t))$. Una vez hecho esto, en cada nuevo estado j, evalúa todos los tokens que le han llegado, y se queda con el mejor de ellos (el que da mayor $\varphi_j(t)$). Entre todos los tokens que en el momento T lleguen al estado final de modelo, se elegirá el de mayor valor, que indica el path de mayor probabilidad.

Entre todos los modelos, el reconocedor de palabras aisladas se quedará con aquel cuyo token elegido sea el de mayor valor de probabilidad.

Esto se puede extender fácilmente a reconocimiento de habla continua, extendiendo el algoritmo anterior para HMMs compuestos:

- Partiendo de un estado inicial, se irán pasando los tokens a los posibles estados conectados.
- Estos estados conectados pueden pertenecer a un mismo HMM, o a HMMs unidos para formar la locución a reconocer.

Debido a lo anterior, es necesario que el reconocedor conozca las posibles transiciones permitidas entre los modelos, que es lo que se conoce como modelo de lenguaje. El modelo de lenguaje indica qué palabras pueden seguir a otras palabras.

Partiendo de un punto de entrada común a la posible red de palabras a reconocer, se irá pasando el token a través de los estados conectados, quedándose con el de mayor valor en cada estado. Con el fin de poder recuperar la locución completa, el token deberá almacenar información de las palabras por las que se va desplazando (cuando pasa del estado de salida de un modelo al estado de entrada de otro). De esta forma, en el momento T el algoritmo se quedará con el mejor token de todos aquellos que se encuentren en el estado final de una palabra, recuperando la lista de palabras por las que ha pasado, que será el resultado del reconocimiento.

Por todo lo dicho hasta ahora, es fundamental que el modelo de lenguaje sea lo más real posible para la aplicación donde se vaya a utilizar, evitando transiciones no deseadas entre palabras (como por ejemplo “el gata” en lugar de “la gata”, lo que se evitaría con un modelo de lenguaje que no permitiera la transición desde “el” a “gata”).

a) HMMs para reconocimiento de habla continua

Hasta ahora se ha hablado de HMMs que representan palabras completas. Para el caso de reconocimiento de habla continua esta elección de modelo no es muy adecuada debido a varias razones[HAH01]:

- Es muy difícil conseguir suficientes locuciones que proporcionen las repeticiones necesarias para un buen entrenamiento de cada HMM.
- Es necesario que cada posible palabra de test esté presente en las locuciones de entrenamiento, haciendo que las tareas de entrenamiento y de reconocimiento no sean independientes.
- Si el vocabulario del reconocedor es muy extenso, el número de modelos no es escalable.
- Los modelos de palabras no capturan los efectos de coarticulación entre palabras. Modelar este efecto se podría plantear únicamente si el vocabulario es muy pequeño (disminuye aún más la escalabilidad).

Para solucionar los problemas anteriores es habitual elegir HMMs que representen unidades acústicas más pequeñas que las palabras, como pueden ser los fonemas. Estos modelos tendrían las siguientes ventajas con respecto a los anteriores:

- Con relativamente pocas locuciones se tendrían suficientes repeticiones de cada modelo.
- No es necesario tener en el entrenamiento todas las palabras de test, sino que es suficiente con que estén todos los fonemas, lo que hace independiente ambas fases del reconocedor.
- El número de modelos es independiente del vocabulario del reconocedor.

Estos modelos presentan un problema fundamental, y es que no caracterizan los efectos de coarticulación, efectos importantes puesto que un fonema puede tener características muy diferentes en función de los fonemas vecinos.

Para tener en cuenta este efecto se tienen los siguientes modelos:

- Sílabas: en este caso el fonema central sí tiene en cuenta el efecto de los vecinos, pero no lo tienen en cuenta los fonemas que se encuentran en los límites de la sílaba.
- Bifonemas y trifonemas: estos modelos representan a un fonema central y la influencia de los fonemas vecinos, ya sea de un solo fonema vecino (para modelar efectos de pausas entre palabras) o de los fonemas inmediatamente anterior y siguiente.

Los modelos usados más extensamente son los bifonemas/trifonemas, puesto que modelan mayor número de características de contexto y además son más escalables que las sílabas. Aún así, siguen presentando debilidades como la dependencia del acento, en función del cual, el mismo trifonema puede presentar diferentes características.

Todo lo explicado anteriormente para reconocimiento automático de habla se puede extender al caso de estos nuevos HMMs. Las principales diferencias son:

- Necesidad de uso de un diccionario que indique los HMMs que forman cada palabra.
- Para el entrenamiento de modelos, se generará un HMM compuesto para la locución de entrada uniendo los HMMs de cada una de las palabras (tarea del diccionario) que indica la trascripción de dicha locución.
- Para el reconocimiento, el token que se pasa entre estados almacenará información de los HMMs por los que pasa, y cuando haya pasado por los que forman una palabra se almacenará esta información (que es la que permitirá obtener la trascripción final).
- Para evitar que en el reconocimiento se elijan siempre las palabras más cortas (que serán las que se encontrarán primero) se pueden fijar unos parámetros para penalizar la inserción de palabras nuevas y para darle mayor o menor peso al modelo de lenguaje. Estos parámetros modificarán el valor de la probabilidad final, que es en lo que se basará el reconocedor para elegir el mejor token (que indica al camino reconocido).

Capítulo 3

Reconocedor automático de habla en castellano

3.1 Resumen

El presente capítulo pretende explicar las distintas etapas que componen el reconocedor automático de habla en castellano desarrollado en el presente proyecto fin de carrera.

Aunque el reconocedor desarrollado pretendía ser un reconocedor de habla continua espontánea, debido a que la base de datos de habla espontánea disponible no ofrecía una cantidad suficiente de datos de entrenamiento para un correcto desarrollo, se decidió implementar un reconocedor de habla continua leída, con una base de datos mayor, y adaptar dicho reconocedor al habla espontánea. Dicha adaptación fue el fruto del proyecto fin de carrera *"Diseño de un reconocedor automático de habla espontánea en castellano"* realizado por *Salvador Alcón Paniagua* [Alc07], del que se van a tomar ciertos desarrollos, puesto que se trató de dos proyectos complementarios.

Por lo tanto, los trabajos realizados en este proyecto fin de carrera utilizan el reconocedor original, implementado para habla continua leída, estudiando la variación de los resultados al introducir en los experimentos información sobre el locutor, ya sea directamente o mediante técnicas de adaptación.

El capítulo actual presenta la implementación básica del reconocedor de habla continua leída, siendo objeto de otros capítulos el estudio de las modificaciones realizadas sobre dicha implementación.

Para el desarrollo realizado se partió de un reconocedor automático de habla continua en inglés, por lo que también se explicarán brevemente las fases que lo componen.

3.2 Reconocedor automático en inglés

En este apartado se va a describir brevemente el reconocedor de habla del que se partió para el desarrollo del reconocedor de habla en castellano objeto del presente proyecto.

Este reconocedor de habla de partida fue desarrollado por el *Departamento de Teoría de la Señal y Comunicaciones* de la *Universidad Carlos III de Madrid*. Se trata de un reconocedor de habla continua en inglés, que utiliza la base de datos *"Wall Street Journal"*.

Para su desarrollo se utilizó lenguaje de scripting bajo Linux, utilizando las herramientas ofrecidas por HTK (Hidden Markov Model Toolkit [YEG+06]).

Los siguientes apartados pretenden ofrecer una visión global de dicho reconocedor, mostrando información sobre las distintas etapas que lo componen, así como sobre la base de datos utilizada. Por último se mencionan las pruebas realizadas sobre el mismo, que supusieron la toma de contacto ante posibles modificaciones que se pueden introducir en el desarrollo y la influencia que toman en los resultados.

3.2.1 Base de datos

El reconocedor que se está estudiando se desarrolló bajo "*The November 1992 ARPA Continuous Speech Recognition Wall Street Journal (CSR-WSJ) Benchmark Test"* [WSJ0], para poder comparar resultados con otros reconocedores que utilicen dicha base de datos.

La base de datos utilizada (*WSJ0*) es una base de datos de habla continua, que almacena lecturas de fragmentos de artículos del periódico "*The Wall Street Journal*" por un total de 123 locutores diferentes, y algunos archivos de habla espontánea. El habla espontánea no será utilizado, por lo que se tratará de un reconocedor de habla continua leída en inglés.

Todas las grabaciones han sido realizadas por dos micrófonos simultáneamente: un Sennheiser HMD410, de tipo micrófono de cabeza ("*head-mounted*"), de forma que se obtenga la mayor calidad de señal posible y con una influencia mínima de ruido exterior, y un segundo micrófono adicional de tipo escritorio ("*desk-mounted*"). La salida producida por cada micrófono para cada frase, muestreada a 16KHz con 16 bits por muestra, se almacena en archivos independientes. Para todos ellos se utiliza un algoritmo de compresión (algoritmo "sorthen", desarrollado por la Universidad de Cambridge) por requerimientos de espacio.

Como datos de entrenamiento, se recomienda elegir entre los conjuntos SI-84, SI-12 y SI-3, siendo SI-84 el utilizado en el reconocedor en cuestión. Se han elegido sólo las grabaciones realizadas con el micrófono Sennheiser, eliminándose de la lista de grabaciones algunas que estaban vacías, obteniendo 7138 locuciones de entrenamiento, de un total de 84 locutores independientes. El que se consideren locutores independientes implica que dichos locutores no van a aparecen en el conjunto de datos de test. Las locuciones del conjunto de entrenamiento elegido se consideran cortas, es decir, con pocas palabras por locución.

Para la evaluación de los resultados, también se presentan distintos conjuntos de datos:

- Datos de habla leída sobre vocabulario de 5000 palabras, independiente del locutor.
- Datos de habla leída sobre vocabulario de 20000 palabras, independiente del locutor.
- Datos de habla espontánea, independiente del locutor.
- Versión leída de los datos de habla espontánea anteriores, independiente del locutor.
- Datos de habla leída sobre vocabulario de 5000 palabras, dependiente del locutor.
- Datos de habla leída sobre vocabulario de 20000 palabras, dependiente del locutor.

En el caso bajo estudio se eligió el primero de los conjuntos, un total de 330 locuciones. En cuanto a las transcripciones, se eligieron las que no ofrecen puntuación verbal.

Por último, mencionar que junto con la base de datos se distribuye un conjunto de modelos de lenguaje, de los cuales se utilizó el bcb05cnp, un modelo de lenguaje de tipo bigram con un vocabulario de 5000 palabras y sin puntuación verbal. Que sea de tipo bigram implica que para una determinada palabra a reconocer, la probabilidad que ofrece el modelo de lenguaje para esa palabra dependerá de la palabra previamente reconocida [WSJ0], [Vic05].

3.2.2 Diccionario

En el sistema bajo estudio se van a utilizar los fonemas como unidades acústicas de reconocimiento. Debido a este hecho se hace necesario el uso de un diccionario que permita conocer la secuencia de modelos acústicos que forman cada palabra (tanto de entrenamiento como de test).

Además, debido a que la base de datos presenta las transcripciones también a nivel de palabras, es necesario este diccionario para obtener dicha información en un formato adecuado para el sistema.

El diccionario que se va a utilizar es la versión 0.6 del diccionario CMU (Carnegie Mellon University Pronouncing Dictionary) [CMUv06]. Es un diccionario de dominio público que utiliza, en su versión no acentuada, un conjunto de 39 fonemas. Se podría utilizar una versión donde se diferenciara el acento de las vocales, pero como se van a emplear los datos sin puntuación verbal de la base de datos, la información de acento no es necesaria.

En la Tabla 1 se muestra el conjunto de fonemas utilizados junto con un ejemplo de palabra que lo contenga y su trascripción correspondiente en fonemas.

Fonema	Ejemplo	Trascripción
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY

EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

Tabla 1.- Conjunto de fonemas utilizados

Por último mencionar que tanto el diccionario como las transcripciones a nivel de fonema se han tenido que adaptar al formato de entrada específico de HTK (Hidden Markov Model Toolkit), el conjunto de herramientas utilizado para el desarrollo del reconocedor.

3.2.3 Tipo de parametrización

La parametrización utilizada consiste en:

- 12 coeficientes MFCC más la log-energía.
- Coeficientes diferenciales de primer orden, velocidades o derivadas primeras de los coeficientes anteriores.
- Coeficientes diferenciales de segundo orden, aceleraciones o derivadas segundas de los primeros.

Estos coeficientes se extraen cada 10 ms utilizando ventanas de análisis de 25 ms, y se tienen un total de 39 coeficientes por cada ventana de parametrización de los datos de entrada.

A estos coeficientes se les aplicará la técnica de normalización de la media cepstral, o técnica CMN. Para cada coeficiente de los 39 que forman esta parametrización, se obtiene el valor medio de ese coeficiente en todas las ventanas que forman la locución, y se resta el valor así obtenido al coeficiente para el que se ha calculado, en cada una de las ventanas. Esto se repite para todas las locuciones a parametrizar.

3.2.4 Topología de los modelos utilizados

Como ya se ha comentado, el fonema será la unidad acústica elegida para este reconocedor. Se tienen un total de 39 fonemas, y cada uno estará modelado por un HMM.

Además de los 39 modelos anteriores (Tabla 1), se añaden otros dos para modelar las pausas entre palabras:

- Modelo ‘sil’: para pausas largas.
- Modelo ‘sp’: para pausas cortas.

La topología elegida para los distintos modelos es la siguiente:

- Modelo de fonema: topología de izquierda a derecha con tres estados emisores, con transiciones al propio estado (sólo si se trata de un estado emisor) o al estado siguiente. Esta topología queda representada en la Figura 13.
- Modelo ‘sil’: no se trata de una topología de izquierda a derecha, ya que al mismo esquema que el utilizado para los modelos de fonemas se le ha añadido una transición que permite ir del último al primer estado emisor. Esta transición se ha añadido para poder modelar mejor los fenómenos que puedan producirse en una pausa de estas características. Esta topología queda representada en la Figura 14.

- Modelo ‘sp’: se modela únicamente con un estado emisor, dada su corta duración en tiempo. Por tener la misma naturaleza que el modelo ‘sil’, este estado emisor estará atado al estado central de dicho modelo (lo que implica que son exactamente iguales y que compartirán datos de entrenamiento). Una peculiaridad de este modelo es que está permitida la transición del estado inicial al final, lo que provoca que se pueda pasar por él sin emitir observación alguna. Esta topología queda representada en la Figura 15.

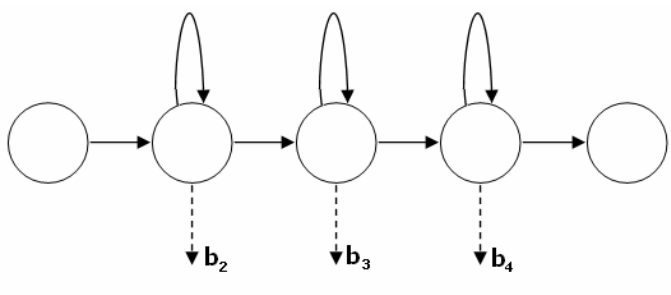

Figura 13.- Topología HMMs fonemas

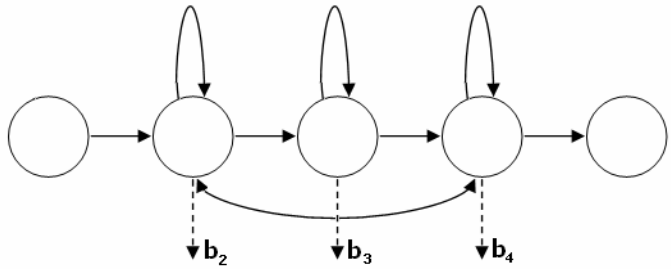

Figura 14.- Topología de HMM ‘sil’

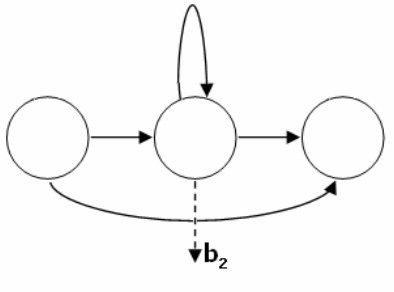

Figura 15.- Topología de HMM ‘sp’

En etapas posteriores del reconocedor se pasará de modelos de fonemas a modelos de trifonemas, manteniéndose la topología de los modelos originales.

3.2.5 Fases del reconocedor

Una vez decididos los datos de entrenamiento y de test, y realizada la parametrización de los mismos, se procede con el entrenamiento de los modelos.

La primera fase del entrenamiento es la que se denomina entrenamiento de monofonos: cada HMM se encargará de modelar un fonema independiente, que en el caso bajo estudio forman un conjunto de 39 patrones acústicos. Aparte de los modelos anteriores se tendrán 2 modelos de silencio ('sil' y 'sp').

Una vez decidida la topología de cada modelo, se pasa a realizar una inicialización de los mismos. Se ha decidido inicializar todos los estados de todos los modelos con el mismo valor, tratándose de una gausiana con una media y una varianza diagonal que se igualarán a la media y varianza de los datos globales (esta última multiplicada por factor menor que 1 para no obtener valores iniciales demasiado elevados). En cuanto a la matriz de transición entre estados, la probabilidad de transición del estado 0 al 1 siempre será de 1, y el resto de transiciones permitidas se inicializará con valor 0.5.

A continuación se comienza con la reestimación de los parámetros de los modelos para adaptarlos a los datos de entrenamiento. En una primera fase se realizan cuatro reestimaciones de los modelos con el algoritmo de Baum-Welch, teniendo en cuenta que el único modelo de silencio que se entrenará será el modelo 'sil', y que éste sólo aparecerá al principio y final de cada locución (para modelar los silencios largos que pueden aparecer en los límites de las grabaciones). En esta primera estimación del modelo 'sil' no aparecen las transiciones entre los estados 2 y 4.

Una vez obtenida una estimación más exacta que la inicial, se pasa a modificar los modelos para que se ajusten más a los datos de entrenamiento. Se añade la transición bidireccional entre los estados 2 y 4 del modelo 'sil' y además se introduce el modelo 'sp'. Para realizar una nueva reestimación se introduce el modelo 'sp' entre palabras, y se vuelve a aplicar el algoritmo de Baum-Welch en otras 4 ocasiones.

Hasta ahora, independientemente de que una palabra pudiera tener varias pronunciaciones, o de que la pausa que le sigue se adapte mejor a un silencio corto o a uno largo, siempre se ha elegido la primera de las pronunciaciones que aparecen en el diccionario para ella, seguida de un silencio corto, excepto en el caso de encontrarnos al final de una frase, donde se utilizará silencio largo. Para elegir de forma más adecuada la pronunciación para cada palabra y el modelo de silencio que le puede seguir, se va a realizar un alineamiento con los datos de

entrenamiento. Para ello, en el diccionario se sustituirá cada pronunciación por otras 2, correspondientes a la inicial acabada en 'sp' y en 'sil' respectivamente.

El alineamiento consiste en realizar un reconocimiento de los datos de entrenamiento utilizando los modelos entrenados hasta el momento, obteniéndose unas nuevas transcripciones que serán más cercanas a la realidad que las originales.

Tras obtener las nuevas transcripciones, se vuelven a estimar los parámetros de los modelos, para adaptarse a los nuevos datos. En esta ocasión sólo se realizan dos ejecuciones del algoritmo de Baum-Welch, entendiendo que serán suficientes para modelar el alineamiento, que debería introducir ligeros cambios.

Hasta aquí el entrenamiento de monofonos. A continuación se pasa a la fase de entrenamiento de trifonemas, donde cada modelo tendrá en cuenta el contexto en el que se encuentra el monofono de partida. Lo primero será obtener una lista con los trifonemas existentes en los datos de entrenamiento, al igual que las transcripciones en función de estos modelos. Cada uno de los HMMs resultantes se inicializará con los datos del HMM del fonema central. A continuación se produce una primera reestimación, mediante dos ejecuciones del algoritmo de Baum-Welch.

Dado que el número de modelos así generado es muy elevado y a que es muy probable que en los datos de entrenamiento no estén presentes todos los trifonemas que se pueden encontrar en los datos de test, se pasa a realizar una agrupación de los modelos (atando diferentes estados de diferentes modelos) mediante un árbol de trifonemas. Si dos modelos tienen atados todos sus estados, se trata de modelos atados. El árbol de trifonemas obtenido, mediante las cuestiones fonéticas desarrolladas para dicho reconocedor, ofrece una reducción del número de estados y, además, modelos que no tienen presencia en los datos de entrenamiento estarán atados gracias a este árbol a otros modelos que sí aparecen en los estos datos, por lo que podrán ser entrenados y, posteriormente, reconocidos.

Una vez obtenido el árbol, cada estado atado se inicializa con las estadísticas de ocupación del estado en el monofono correspondiente, y se realiza una reestimación de los modelos mediante 4 ejecuciones del algoritmo de Baum-Welch.

A partir de este punto, las siguientes fases del entrenamiento se encargan de aumentar el número de gausianas que componen cada estado. Se pasa primero de 1 a 2 gausianas por estado, reestimando los modelos mediante 4 ejecuciones del algoritmo de Baum-Welch. A continuación se pasa de 2 a 4, y después de 4 a 8, siempre incluyendo un paso de reestimación después de duplicar el número de gausianas. Por último, y sólo para los modelos de silencio, se aumenta de 8 a 16 gausianas.

En este punto finaliza el entrenamiento de los modelos presentes en el reconocedor, estando disponibles para el reconocimiento.

En la fase de test, se aplica el algoritmo de Viterbi (en el caso bajo estudio, por tratarse de HTK, es una modificación de dicho algoritmo conocido como algoritmo de paso de token (Token Passing)) para realizar el reconocimiento de las frases consideradas. En esta etapa, es necesario fijar el valor de dos parámetros importantes: WIP (penalización por inserción de palabra) y LM (peso del modelo de lenguaje).

Por último, mencionar que HTK presenta una herramienta que permite analizar los resultados de reconocimiento comparándolos con las transcripciones exactas de los datos de test, de forma que se puedan presentar resultados fácilmente entendibles y comparables.

3.2.6 Experimentos realizados

Sobre el reconocedor anteriormente descrito se han realizado una serie de experimentos con el fin principal de obtener conocimiento sobre la manipulación de los scripts que lo forman y sobre el uso de las herramientas de HTK necesarias. Además, las pruebas realizadas muestran la variación de los resultados de reconocimiento en función de la configuración aplicada a algunos parámetros.

Algunas de las conclusiones obtenidas son:

- El aumento de la frecuencia de muestreo influye positivamente en los datos de reconocimiento. Eran resultados esperados, ya que al aumentar la frecuencia de muestreo, los datos obtenidos tienen más información sobre la señal real.
- La aplicación de la técnica CMN en la parametrización ofrece mejores resultados que si no la utilizamos. Esta técnica intenta eliminar, o al menos, disminuir la influencia de los micrófonos, aportando buenos resultados.
- En cuanto a la eliminación de los coeficientes de aceleración de la parametrización, esto hace que empeoren en gran medida los resultados, lo que indica que aunque aumente bastante la dimensión de los vectores de parámetros, y por lo tanto el tiempo de procesado, es una buena opción utilizarlos.
- La mejora producida al permitir los 2 modelos de silencio entre palabras, en lugar de utilizar sólo el modelo de silencio corto, justifican el aumento en la complejidad.
- Se ha comprobado la importancia del número de reestimaciones realizadas tras cada modificación de los parámetros de los modelos.

- También se ha comprobado la gran mejora producida al pasar de modelos de monofonos a modelos de trifonemas, igual que al ir mejorando los modelos de trifonemas (tanto con el árbol de agrupamiento como con el aumento del número de gausianas). Sólo se ha producido un pequeño empeoramiento en los resultados al pasar de 8 a 16 gausianas para los silencios, lo que indica que los modelos podrían estar adaptándose demasiado a los datos de entrenamiento, y como el conjunto de test está formado por locutores independientes, la mejora de prestaciones no es tan considerable como se esperaba.
- Por último se observó el efecto de la penalización por inserción de palabra y del peso del modelo de lenguaje. A medida que el valor del primero toma valores más negativos, lo que indica que introducir una palabra nueva en el reconocimiento tiene una penalización mayor, se observa como el número de inserciones disminuye, pero también aumenta el número de eliminaciones. En cuanto al peso del modelo de lenguaje, a medida que aumenta se le da mayor importancia al modelo de lenguaje. Si es demasiado elevado puede llegar a anular la importancia de la probabilidad acústica del modelo. Este hecho se ve reflejado en un aumento de las eliminaciones a medida que aumenta este factor. Debido a que el número de inserciones y de eliminaciones varían de forma inversa al tocar estos parámetros, lo óptimo sería elegir un valor para ambos parámetros donde el número de inserciones y de eliminaciones se igualen, lo que aporta los mejores valores para la tasa de acierto.

3.3 Reconocedor automático en castellano

En este apartado se van a describir los desarrollos llevados a cabo para la implementación del reconocedor de habla continua leída en castellano que va a usarse en el estudio realizado en el presente proyecto fin de carrera.

Como ya se ha comentado, el objetivo final es implementar un reconocedor de habla espontánea, y utilizar distintas técnicas para, aplicando información sobre los locutores, observar la variación en los resultados de test y sacar conclusiones sobre las técnicas que proporcionan las mejoras más significativas. Como la base de datos de habla espontánea disponible (TC-STAR) no tiene los suficientes datos de entrenamiento como para un desarrollo directo del reconocedor de habla espontánea, se decidió utilizar una base de datos de habla leída en castellano (MICROAES), con una cantidad mucho más extensa de datos de entrenamiento, y adaptar los modelos así obtenidos con los datos de habla espontánea. En este proyecto se

realizará el estudio planteado anteriormente, pero utilizando el reconocedor previo a la adaptación al habla espontánea.

Al igual que el reconocedor en inglés ya estudiado, el reconocedor en castellano también está desarrollado bajo Linux con un lenguaje de scripting, mediante el que se utilizarán las herramientas ofrecidas por HTK (Hidden Markov Model Toolkit) para la realización de reconocedores de habla basados en modelos ocultos de Markov (HMM).

Los siguientes apartados pretenden ofrecer una visión específica sobre el desarrollo que se llevó a cabo, mostrando información sobre las distintas etapas que lo componen, así como sobre la base de datos utilizada. Por último se mencionará la prueba básica realizada sobre el mismo, prueba que se utilizará como punto de referencia para comprobar los resultados obtenidos al introducir información sobre los locutores en la generación de los modelos.

3.3.1 Base de datos

MICROAES (ATLAS Spanish Microphone DataBase) [MIC04] será la base de datos utilizada para el desarrollo del reconocedor de habla continua, que fue creada por una empresa española dedicada a las tecnologías del habla, ATLAS (Applied Technologies on Language and Speech).

Se trata de una base de datos de habla leída, para la que se han elegido un total de 450 párrafos distintos (agrupados en 30 grupos de 15 párrafos cada uno) obtenidos de diferentes periódicos y libros, con temas variados, para obtener un vocabulario lo más extenso posible.

Para que la base de datos fuera lo más balanceada posible, además de cuidar el contenido de las locuciones, también se cuidó la elección de los locutores, que fueron seleccionados en función del género, así como de las variedades dialectales, y de la edad (Tabla 2 y Tabla 3).

Dialecto	Masculinos	Femeninos	Total (nº/%)
Centro	35	41	76 (25.33%)
Este	29	30	59 (19.67%)
Norte	21	23	44 (14.67%)
Noroeste	22	24	46 (15.33%)
Sur	40	35	75 (25.00%)
Total	147	153	300 (100%)

Tabla 2.- Información locutores en función de variedad dialectal y género.

Edad	Masculinos	Femeninos	Total (nº/%)
< 15	6	8	14 (4367%)
16-30	63	62	125 (41.67%)
31-45	31	36	67 (22.23%)
46-60	34	34	68 (22.67%)
> 60	13	13	26 (8.67%)
Total	147	153	300 (100%)

Tabla 3.- Información locutores en función de edad y género.

La base de datos cuenta con un total de 300 locutores diferentes. Por cada uno de ellos se tienen 15 frases de entrenamiento diferentes más 2 frases (las mismas para todos los locutores) de test, grabadas por 4 micrófonos simultáneamente.

De los 4 micrófonos, dos de ellos están colocados a muy corta distancia del locutor (de forma que se obtengan grabaciones de voz lo más limpia posible, con influencia mínima de ruidos externos), el tercero a media distancia y el cuarto a gran distancia (de 2 a 3 metros).

La salida producida por cada micrófono para cada frase, muestreada a 16KHz con 16 bits por muestra, se almacena en archivos independientes en formato WAV y sin cabeceras. No se aplica ningún algoritmo de compresión.

La base de datos cuenta con 18000 locuciones de entrenamiento (teniendo en cuenta los 4 micrófonos) con 7410 palabras, y con 2400 locuciones de test con 145 palabras diferentes (número de palabras reducidas debido a que sólo existen 2 frases de test diferentes).

Para poder utilizarla, es necesario conocer la estructura de almacenamiento de la información en la base de datos, y así poder crear las listas de entrenamiento y de test. Existe un directorio inicial, organizado en 30 directorios diferentes, conteniendo cada uno de ellos la información sobre 10 locutores, estando la información de cada locutor almacenada en un directorio individual.

En cuanto a la información concreta de la locución, ésta está incluida en el nombre del archivo de audio que la almacena. Este nombre estará compuesto por 11 caracteres (más otros 3 indicando la extensión del archivo), proporcionando la siguiente información:

- Los cinco primeros caracteres indican el locutor al que pertenece (‘ma000’ – ‘ma299’).
- Tres caracteres más para la sesión. El primero de ellos indica si se trata de una sesión de entrenamiento (‘s’) o de test (‘x’), y los otros dos caracteres permiten

identificar la sesión dentro de su propio grupo (01 – 15 para sesiones de entrenamiento, 01 – 02 para sesiones de test).

- Los 4 últimos indican cuál de los 4 micrófonos se ha utilizado para su grabación ('es0' – 'es3').

Además de las locuciones, la base de datos ofrece:

- Las transcripciones correspondientes, sin puntuación verbal. Estas transcripciones, además de las palabras, ofrecen información no verbal, señalizando los silencios, dudas o ruidos realizados por el locutor, y ruidos que se puedan producir en el entorno al realizar la grabación.
- Un diccionario con más de 7400 palabras, donde la pronunciación de cada palabra ha sido obtenido por el transcriptor fonético SAGA (Spanish Automatic Graphemes to Allophones Transcriber), que utiliza la notación fonética SAMPA (Speech Assessment Methods Phonetic Alphabet). Aunque es un diccionario muy completo para esta base de datos, puesto que contendrá todas las palabras presentes en los datos de entrenamiento y test, no es suficiente para lo que se pretende con el reconocedor desarrollado, debido a que tendrá que adaptarse con otras bases de datos. Por lo tanto, será necesario el desarrollo de un nuevo diccionario.

Para los desarrollos realizados en este proyecto fin de carrera se utilizaron como grupo de locuciones de entrenamiento las grabaciones realizadas con los micrófonos que se encuentran más cerca del locutor ('es1' y 'es2') de las 15 frases que se consideran de entrenamiento de los 300 locutores, y como locuciones de test las grabaciones realizadas con esos mismos micrófonos sobre las 2 frases de test de cada locutor.

3.3.2 Diccionario

El reconocedor de habla continua en castellano también utilizará los fonemas como unidades acústicas de reconocimiento. Debido a este hecho se hace necesario el uso de un diccionario que permita conocer la secuencia de modelos acústicos que forman cada palabra (tanto de entrenamiento como de test).

Además, debido a que la base de datos también presenta las transcripciones a nivel de palabra, es necesario este diccionario para obtener dicha información en un formato adecuado para el sistema.

Como ya se ha mencionado, la base de datos ofrece un diccionario de más de 7400 palabras, que contiene todas las palabras presentes en las frases de entrenamiento y de test.

Para crear este diccionario se usó el transcriptor fonético SAGA, que utiliza la notación fonética SAMPA, descrita en la Tabla 4.

Fonema	Ejemplo	Trascripción
p	perro	**p** ‘e rr o
B	comba	K ‘o m **b** a
t	toro	**t** ‘o r o
d	caldo	k ‘a l **d** o
k	kiosko	**k** j ‘o s **k** o
g	tongo	t ‘o N **g** o
m	arma	‘a r **m** a
n	cono	k ‘o **n** o
N	anca	‘a **N** k a
J	uña	‘u **J** a
tS	chelo	**tS** ‘e l o
f	cofia	k ‘o **f** j a
T	celo	**T** ‘e l o
s	casa	k ‘a **s** a
z	rasgo	rr ‘a **z** G o
jj	yunque	**jj** ‘u n k e
x	genio	**x** ‘e n j o
l	lote	**l** ‘o t e
L	tallo	t ‘a **L** o
rr	carro	k ‘a **rr** o
j	armario	a r m ‘a r **j** o
w	cigüeña	T i G **w** ‘e J a
B	labio	l ‘a **B** j o
D	codo	K ‘o **D** o
G	lago	l ‘a **G** o
r	arpa	‘a **r** p a
a	honra	‘o n rr **a**
e	queso	k ‘**e** s o
i	tizne	t ‘**i** T n e
o	calvo	k ‘a l B **o**
u	lujo	l ‘**u** x o

Tabla 4.- Conjunto de fonemas utilizados

Como se puede ver en la Tabla 4, esta notación muestra un total de 31 fonemas. Sin embargo, el transcriptor ofrece información de las vocales con acento prosódico de cada palabra, por lo que se decidió añadir los vocales acentuadas al conjunto de fonemas, obteniéndose un total de 36 unidades acústicas diferentes, que quedan representados en la Tabla 5.

p	B	T	d	k	g	m	n	N	J	tS	f
T	S	Z	jj	x	l	L	rr	j	w	B	D
G	R	A	‘a	e	‘e	i	‘i	o	‘o	u	‘u

Tabla 5.- conjunto de unidades acústicas final

El diccionario ofrecido por MICROAES debe ser adaptado para su utilización en el reconocedor de habla continua que se va a desarrollar. En concreto HTK presenta las siguientes limitaciones:

- Organización del diccionario: dos columnas, sin título, donde la primera columna está formada por las palabras del diccionario, tal y como aparecerán en las transcripciones, y la segunda columna será la transcripción según los modelos utilizados por el reconocedor.
- Restricción de caracteres: los caracteres “**á, é, í, ó, ú, ñ, ü”** no son admitidos por HTK.

Para adaptar el diccionario al anterior formato fue necesario:

- Eliminar una columna intermedia que ofrecía información sobre la frecuencia de aparición de las distintas palabras del diccionario en la base de datos.
- Cambiar los caracteres prohibidos en HTK por otros que se pudieran interpretar fácilmente. En concreto se hicieron los cambios que aparecen en la Tabla 6. Notar que algunas de estas sustituciones (**ñ** y **ü**) sólo pueden realizarse porque todas las palabras del diccionario están en minúsculas (sino, sería imposible diferenciar una **ñ** de una **N**, o una **ü** de una **U**).

Caracter no permitido	Sustitución
á	a1
é	e1
í	i1
ó	o1
ú	u1
ñ	N
ü	U

Tabla 6.- cambio de caracteres no permitidos por HTK.

Aunque el transcriptor fonético SAGA muestra información sobre el acento prosódico en las transcripciones generadas, esa información fue eliminada del diccionario de MICROAES. Como el propósito original era tener en cuenta esa información para modelar mejor la señal de voz, se decidió realizar de nuevo la trascripción fonética de este diccionario. Para esto se siguieron los siguientes pasos:

- Obtención de la lista de palabras en minúsculas.
- Adaptación al formato de entrada del transcriptor (según muestra la Tabla 7).
- Realización de la transcripción fonética mediante el transcriptor fonético SAGA.
- Adaptación de la salida del transcriptor al formato esperado por HTK, donde sólo habrá que transformar las vocales acentuadas **'a, 'e, 'i, 'o** y **'u** de las transcripciones, por **a1, e1, i1, o1** y **u1**.
- Unir las dos listas para formar el diccionario final.

Caracter no permitido	Sustitución
á	'a
é	'e
í	'i
ó	'o
ú	'u
ñ	~n
ü	~u
¿	'?
¡	'!

Tabla 7.- cambio de caracteres no permitidos por SAGA

El diccionario ofrecido por la base de datos MICROAES, adaptado a los requisitos del reconocedor, es suficiente para los datos de entrenamiento y test del reconocedor de habla continua leída. Sin embargo, cuando este reconocedor se adapta a las características de habla espontánea, para lo que utiliza una base de datos diferente (en concreto, TC-STAR), puede que el diccionario anterior no sea suficiente. Por este motivo se desarrolló un nuevo diccionario, más general y mucho más extenso que el anterior.

El nuevo diccionario parte de una lista de palabras proporcionadas por el grupo de trabajo COES, que pertenece al Departamento de Arquitectura y Tecnología de Sistemas Informáticos de la Universidad Politécnica de Madrid [COE05]. En concreto esta lista está formada por 50.000 palabras inicialmente, a las que se le han añadido conjugaciones de verbos, palabras compuestas (prefijos y sufijos) y combinaciones de palabras, originando aproximadamente unas 700000 palabras.

Con esta lista de palabras se formará el nuevo diccionario, para lo que fue necesario:

- Eliminar las abreviaturas existentes, que no tienen sentido en el ámbito de trabajo del diccionario.
- Adaptar la lista al formato de entrada del transcriptor fonético SAGA.
- Trascripción.
- Adaptar la lista de palabras y las transcripciones al formato esperado por HTK (Tabla 6).
- Unir la información anterior en dos columnas para formar finalmente el diccionario para HTK.

Después de obtener el diccionario, es necesario adaptar las transcripciones de las locuciones presentes en la base de datos. Estas transcripciones indican las palabras que forman cada locución. Son necesarias tanto en la fase de entrenamiento para obtener el conjunto de modelos que forman la locución, como en la fase de test para poder obtener medidas sobre los resultados. Por lo tanto, habrá que adaptar estas transcripciones al formato esperado por HTK.

Además de la anterior, es necesaria una segunda adaptación, ya que las palabras que forman la primera columna del diccionario y las palabras que forman las transcripciones deben compartir el mismo formato, con las limitaciones mostradas en la Tabla 6. Esto es necesario tanto para entrenamiento, puesto que se necesitará buscar en el diccionario la trascripción fonética de cada palabra de la locución para obtener los HMMs correspondientes, como para reconocimiento, donde reconocidos un conjunto de HMMs se ofrecerá como salida de reconocimiento las palabras correspondientes a esos modelos, que se compararán con las indicadas por la transcripción real de la locución para mostrar los resultados.

3.3.3 Tipo de parametrización

En cuanto al tratamiento de los archivos de voz, el primer paso en todo reconocedor de habla es la parametrización. Como ya se ha comentado, este proceso consiste fundamentalmente en obtener las características más significativas de la señal de voz, de forma que se tenga sólo la información que interviene en el reconocimiento. Además, al tratarse de una extracción de información, se reduce el espacio necesario para el almacenamiento de los datos, tanto de entrenamiento como de test.

Existe una herramienta en HTK que permite parametrizar los archivos de voz. Los parámetros resultantes dependerán de la configuración que se le indique a esta herramienta. En

concreto, para el reconocedor bajo estudio, se está indicando la siguiente información para realizar la parametrización:

- Los archivos de voz se presentan almacenados en formato WAV, sin cabeceras y muestreados a 16KHz con 16 bits por muestra.
- La obtención de cada segmento a parametrizar debe realizarse mediante ventanas Hamming con un tamaño de 25 mseg., siendo la separación entre ventanas consecutivas de 10 mseg.
- Se va a realizar un filtrado de preénfasis, con el objetivo de aumentar la amplitud de las altas frecuencias para corregir, en parte, la atenuación que el proceso de generación de voz introduce en ellas. El coeficiente de preénfasis se fija en 0.97.
- Para cada ventana, el proceso de parametrización obtendrá un total de 39 coeficientes, dividiéndose en 12 coeficientes MFCC (obtenidos a partir de un banco de 40 filtros en la escales Mel) más la log-energía, y las deltas y aceleraciones de los coeficientes anteriores.
- La energía se debe almacenar normalizada para cada locución, lo que implica restar al valor de energía de cada ventana el valor máximo que toma en la locución, sumándole 1. Además hay que tener en cuenta que se fija un rango máximo de variación de la energía, 50 dB en este caso, impidiendo que se produzcan valores demasiado pequeños. Por último, se indica que hay que aplicar un factor de escalado igual a 0.1.
- La energía debe calcularse antes de cualquier enventanado o filtro de preénfasis.
- Finalmente, a los coeficientes se le aplica la técnica CMN para reducir su dependencia de los micrófonos. Para ello, se calcula el valor medio de cada uno de los 12 MFCCs y de la log-energía a lo largo de cada locución, eliminando la media de cada coeficiente de su valor real en cada ventana.

3.3.4 Topología de los modelos utilizados

Al igual que en el reconocedor automático de habla en inglés del que se partió, el reconocedor en castellano utilizará los fonemas como unidades acústicas básicas a modelar. Por lo tanto se tendrá un HMM para cada uno de los 31 fonemas del castellano. Además de esto, se han añadido 5 HMMs más que modelarán las vocales acentuadas, ya que se considera que éstas presentan características marcadas que las diferencian de las no acentuadas, permitiendo un modelado por separado, y que el poder distinguir entre los dos grupos permitirá mejoras en el reconocimiento.

Además de los 36 HMMs anteriores (Tabla 5) que pueden asociarse a fonemas, se van a añadir otros 3:

- Dos modelos para modelar las pausas entre palabras.
  - 'sil': para pausas largas.
  - 'sp': para pausas cortas.
- Un modelo basura para modelar ruidos, dudas o palabras no acabadas emitidas por los locutores, considerándose como información no útil para el reconocimiento.

El modelo basura aunque no es demasiado importante en el desarrollo de un sistema de reconocimiento de habla continua leída, se considera fundamental cuando se trabaja con habla espontánea, causa que motivó su utilización. Aunque la base de datos diferencia entre distintos casos que se pueden considerar como información no útil, se decidió agrupar todos ellos en un sólo modelo, principalmente porque el número de repeticiones, si se consideraban individualmente, no era suficiente para un buen entrenamiento.

La topología elegida para los distintos modelos es la siguiente:

- Modelo de fonema: topología de izquierda a derecha con tres estados emisores, con transiciones al propio estado (si se trata de estado emisor) o al estado siguiente. Esta topología queda representada en la Figura 13.
- Modelo 'sil': ya no se trata de una topología de izquierda a derecha, ya que al mismo esquema que el utilizado para los modelos de fonemas se le ha añadido una transición que permite la conexión directa entre el primer y último estado emisor, tratándose de una transición bidireccional. Esta transición se ha añadido para poder modelar mejor los fenómenos que puedan producirse en una pausa de estas características. Esta topología queda representada en la Figura 14.
- Modelo basura: utiliza la misma topología que el modelo 'sil'. La transición hacia atrás da un grado más de libertad a este modelo para ajustarse mejor a las muestras que intenta modelar y que no presentan la evolución habitual de las muestras de habla.
- Modelo 'sp': se modela únicamente con un estado emisor, dada su corta duración en tiempo. Por tener la misma naturaleza que el modelo 'sil', este estado emisor estará atado al estado central de dicho modelo (lo que implica que son exactamente iguales y que compartirán datos de entrenamiento). Una peculiaridad de este modelo es que está permitida la transición del estado inicial al final, lo que provoca que puede atravesarse sin que produzca ninguna observación. Esta topología queda representada en la Figura 15.

En etapas posteriores del reconocedor se pasará de modelos de fonemas a modelos de trifonemas, manteniéndose la misma topología de los modelos originales.

3.3.5 Fases del reconocedor

En este apartado se pretende explicar detalladamente las distintas fases del reconocedor desarrollado.

a) Inicialización de los modelos

Lo primero que se tuvo que elegir fue la separación de los datos disponibles en datos de entrenamiento y datos de test. La base de datos MICROAES sugiere una separación específica, y se decidió utilizarla. Se tendrán 300 locutores de entrenamiento, y esos mismos locutores en la fase de test, por lo que no se puede considerar que los resultados obtenidos con esta división sean independientes del locutor, condición que no se considera necesaria para el objetivo del proyecto.

Para cada locutor se dispone de 15 frases de entrenamiento diferentes. Aunque cada frase ha sido grabada por 4 micrófonos distintos, sólo se han elegido las locuciones obtenidas por los micrófonos más cercanos al locutor, y por lo tanto, aquellas muestras de más calidad, con influencia mínima de ruido ambiente.

En cuanto a los datos de test, se disponen de 2 frases por cada locutor (las mismas para todos los locutores) eligiéndose, de nuevo, sólo las locuciones grabadas por los dos micrófonos más próximos.

Por lo tanto, se tendrán un total de 9000 locuciones de entrenamiento y 1200 de test, para un conjunto de 300 locutores, apareciendo cada uno de ellos tanto en los datos de entrenamiento como de test.

Una vez elegidos los datos que se van a utilizar y partiendo de la parametrización de dichas muestras, se procede a la inicialización de los modelos.

Para inicializar los distintos HMMs que componen el conjunto de modelos del reconocedor bajo estudio se ha decidido dar los mismos valores iniciales a cada estado emisor de cada HMM. Cada uno de estos estados se modelará con una mezcla de gausianas, partiendo de una mezcla simple compuesta por una única gausiana.

Dado que cada muestra parametrizada contiene 39 componentes, la media de la gausiana será un vector de longitud 39 y la varianza será una matriz longitudinal de dimensión 39x39, aunque hay que tener en cuenta que se utilizan matrices diagonales. Se ha decidido utilizar la

media y varianza globales de los datos de entrenamiento como valor inicial de la gausiana que compone cada estado de cada HMM.

En cuanto a la matriz que indica las posibles transiciones entre estados, se inicializará de forma que sólo se puedan dar las transiciones propias de cada modelo, y las posibles transiciones desde cada estado sean equiprobables.

En esta primera inicialización el modelo basura y el modelo de silencio 'sil' no tienen la transición entre los estados 2 y 4, y el estado central de ambos modelos está atado, lo que quiere decir que serán iguales y que compartirán datos de entrenamiento. El estado emisor del modelo de silencio 'sp' también está atado al estado central de los modelos anteriores.

b) Entrenamiento de monofonos

En esta fase se va a realizar el entrenamiento de los modelos descritos hasta el momento: 3 modelos especiales ('sil', 'sp' y modelo basura) y 36 modelos que representan los distintos fonemas considerados (tratando las vocales acentuadas como modelos de este tipo). En todos ellos se modelará cada estado con una única gausiana, procediéndose a aumentar el número de gausianas en etapas posteriores del entrenamiento.

Como punto de partida se tienen todos los modelos inicializados con la media y varianza global de los datos de entrenamiento.

En el entrenamiento de HMMs, se producen mejores resultados si se realizan pequeños cambios en los modelos y se adaptan los modelos a cada cambio, que si se entrenan desde el principio los modelos que se quieren obtener finalmente. Debido a esto, el entrenamiento actual se divide en varias fases, realizándose en cada una de ellas una reestimación de los modelos tras aplicarle los cambios correspondientes. Esta reestimación se realiza mediante el algoritmo de Baum-Welch, utilizando más o menos ejecuciones del mismo en función de lo significativo que pueda ser el cambio aplicado en los modelos. Se marca un número fijo de iteraciones del algoritmo en lugar de realizar tantas como sean necesarias para llegar a la convergencia por varias razones:

- Conseguir la convergencia puede ser demasiado costoso en cuanto a tiempo de ejecución.
- Si se llega a esta convergencia se corre el riesgo de que los HMMs estén sobreadaptados a los datos de entrenamiento, produciendo malos resultados de reconocimiento.

El número de veces que se aplique el algoritmo debe ser suficientemente alto como para que los parámetros de los modelos se adapten a los datos, y lo suficientemente bajo como para que no se produzca esa sobreadaptación a los datos de entrenamiento.

La primera fase de este entrenamiento será la reestimación de los modelos inicializados. Los datos para este entrenamiento serán las transcripciones de las locuciones considerando que no existe pausa entre palabras, y que las únicas pausas posibles se introducen al principio y final de la grabación, modelándose dichas pausas con el modelo de silencio 'sil'. El modelo basura estará presente en las transcripciones, puesto que la base de datos indica la posición de la etiqueta correspondiente. Sin embargo, en este primer entrenamiento, el modelo de silencio 'sp' no va a aparecer en las transcripciones, por lo que se considera que no existen muestras que lo representen para poder reestimarlo.

Debido a que las medias y varianzas de todos los estados de todos los modelos son inicializados con el mismo valor se considera que esta primera adaptación a los datos tendrá un impacto elevado, por lo que el algoritmo de Baum-Welch se ejecutará en 4 ocasiones.

A continuación se realizan los siguientes cambios a los modelos:

- Se modifica el modelo de silencio 'sil' añadiendo una transición bidireccional entre los estados 2 y 4 de dicho modelo, con una probabilidad de transición en cada sentido de 0.2.
- Se modifica el modelo basura de la misma forma que el modelo 'sil'.
- Se introduce el modelo de silencio 'sp' en la lista de modelos a entrenar.

Una vez realizados estos cambios se procede a reestimar los modelos. Para poder entrenar el modelo 'sp' es necesario modificar las transcripciones de las locuciones, puesto que hasta ahora este modelo no aparecía en ellas. Se introduce el nuevo modelo de silencio como pausa entre palabras, manteniendo el modelo 'sil' como pausa larga al principio y final de cada locución.

Ya que los cambios introducidos son de importancia, se ejecutará el algoritmo de Baum-Welch en 4 ocasiones para adaptar los nuevos modelos a los datos de entrenamiento.

Como se ha podido observar la inclusión de las pausas en las transcripciones se ha realizado de forma completamente arbitraria, atendiendo únicamente a lo que se espera (pausas largas al principio y final de grabación, y pausas cortas entre palabras) pero sin tener referencia alguna de lo que está sucediendo en realidad. Con el fin de conseguir unas transcripciones más representativas, se va a realizar un proceso conocido como alineamiento.

El alineamiento consiste es un reconocimiento, mediante el algoritmo de Viterbi, con los modelos entrenados hasta el momento, donde se deberá decidir entre las opciones disponibles para cada palabra. Si una palabra tiene diferentes pronunciaciones, normalmente se elige una pronunciación al azar para las primeras fases del entrenamiento. Después, mediante el alineamiento, se evalúan las distintas pronunciaciones de dicha palabra, eligiendo la que produzca una verosimilitud mayor.

En este caso, además de realizar una adaptación a las diferentes pronunciaciones de una misma palabra, se pretende elegir el modelo más adecuado a colocar entre palabras (modelo basura, modelo 'sil' o modelo 'sp'). Esto se puede hacer modificando el diccionario, de forma que por cada pronunciación de una determinada palabra ahora se van a tener tres diferentes. La diferencia entre cada una de ellas estriba en que al final se añadirá uno de los tres modelos que se pueden utilizar para la separación entre palabras.

Una vez realizado el alineamiento con el nuevo diccionario, las transcripciones habrán sido modificadas, presentando entre palabras el modelo que más se adapte a las locuciones reales de entrenamiento.

Se hace necesaria una reestimación de los modelos con las nuevas transcripciones. Como no se esperan unos cambios de gran envergadura tras este proceso, sólo se realizan 2 ejecuciones del algoritmo de Baum-Welch.

Con esto finaliza este primer entrenamiento de los modelos. A lo largo de esta fase se han considerado modelos de fonemas, donde no se tiene en cuenta el contexto de los mismos. En la siguiente fase del entrenamiento se realizan las modificaciones necesarias para entrenar otro tipo de modelos que tenga en cuenta la influencia de los fonemas contiguos, ya que las características de un fonema pueden verse muy influenciadas por los fonemas anterior y/o siguiente.

c) Entrenamiento de trifonemas

En esta fase se van a modificar los modelos entrenados en la fase anterior para hacerlos dependientes del contexto en el que se encuentran, pasando de HMMs asociados a fonemas a HMMs asociados a trifonemas. Además, una vez se hayan entrenado lo suficiente, se irá aumentando el número de gausianas que componen cada estado, con el fin de proporcionar una mayor capacidad de modelado.

En esta fase del entrenamiento se parte de un conjunto de HMMs que representan fonemas (más los 3 HMMs que modelan las transiciones entre palabras). El primer paso para generar los

trifonemas es identificar los modelos de trifonemas presentes en los datos de entrenamiento, para lo que es necesario modificar las transcripciones de las locuciones resultantes del alineamiento, sustituyendo cada fonema por el correspondiente trifonema en función del contexto del mismo. Las reglas para realizar esta conversión son las siguientes:

- Un fonema 'f' pasará a ser el trifonema 'i'-'f'+'d', siendo 'i' el fonema que se encuentra a su izquierda y 'd' el fonema que se encuentra a su derecha.
- Si en las transcripciones de las locuciones se encuentra el modelo 'sp' o el modelo basura, estos no generarán ningún trifonema, ni como fonema central ni como contexto.
- En el caso del modelo 'sil', sólo podrá formar parte del contexto de los trifonemas que se generan a partir de otros fonemas, nunca se generará un trifonema partiendo del modelo 'sil' como fonema central.

Un ejemplo de la transformación sufrida por una trascripción se representa en Figura 16:

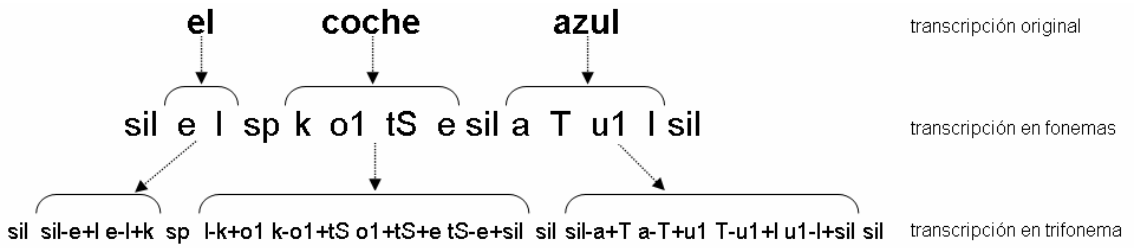

Figura 16.- modificación de trascripción al pasar a trifonemas

A continuación se proporcionará un valor inicial a los nuevos modelos, que compartirán la misma topología que los anteriores. Los modelos de trifonemas resultantes de un mismo fonema central se inicializarán con los valores del HMM que representa a dicho fonema. Una particularidad de los trifonemas formados a partir del mismo fonema es que se atan las matrices de transición, lo que quiere decir que todos estos trifonemas contribuyen a su modificación, y que todos ellos comparten siempre los mismos valores.

Para adaptar los nuevos modelos a los datos de entrenamiento se realizarán dos ejecuciones del algoritmo de Baum-Welch. De esta forma, se han adaptado los modelos de fonemas a su contexto.

Como ya se ha comentado, los trifonemas que se han entrenado hasta el momento son únicamente los que aparecen en los datos de entrenamiento. Por lo tanto estos modelos no son adecuados para realizar el reconocimiento, ya que no se tiene seguridad de entrenar todos los posibles modelos que puedan aparecer en la fase de test.

Debido a que el número de trifonemas es demasiado elevado como para tener suficientes realizaciones de cada uno en la fase de entrenamiento, es habitual el uso de árboles binarios de agrupamiento para poder utilizar este tipo de modelos.

Los árboles de agrupamiento indican grupos de modelos con características similares. Si uno de estos grupos está formado por *N* modelos diferentes, en lugar de necesitar entrenar *N* HMMs sólo se entrenará uno, utilizando para ello los datos disponibles de los *N* modelos que forman el grupo. De esta forma, se tendrán HMMs bien entrenados que serán compartidos por varios trifonemas, que pueden darse o no en los datos de entrenamiento.

Un árbol binario de agrupamiento parte de un conjunto de modelos inicial, al que se realizan una serie de cuestiones. Cada una de estas cuestiones divide en dos el conjunto de modelos iniciales. Esta división debería aumentar la log-verosimilitud de los datos (todo se realiza sobre los datos de entrenamiento), ya que se obtendrán grupos de modelos que comparten más características. Se elegirá la cuestión que mayor aumento de la log-verosimilitud produzca, y sobre los grupos resultantes de la división anterior se vuelve a ejecutar el mismo proceso de nuevo. El criterio de parada viene fijado por tres factores diferentes:

- La aplicación de todas las cuestiones disponibles.
- El aumento de la log-verosimilitud es inferior a un umbral marcado. Si este hecho se produce, no se realiza esta última división.
- El número de muestras disponibles en los datos de entrenamiento para un determinado grupo es inferior a un segundo umbral marcado. Llegado este punto, no se realiza la última división.

En el caso del reconocedor desarrollado, lo que se pretende es obtener grupos de modelos que presenten características acústicas semejantes. Por lo tanto, la primera división de los datos será por grupos de trifonemas que compartan el fonema del estado central, ya que se supone que todos ellos comparten una forma de onda semejante que se verá ligeramente modificada por el contexto.

A cada uno de estos grupos será al que se va a aplicar el proceso de agrupamiento del árbol binario, aplicando una serie de cuestiones fonéticas, que irán clasificando los distintos grupos en función del tipo de fonema que se presente a su derecha y a su izquierda. Si se tienen dos trifonemas diferentes del mismo fonema central, y tanto el fonema que se encuentra a la izquierda de un modelo como del otro pertenecen al mismo tipo de fonemas, parece lógico pensar que la modificación que pueden presentar ambos contextos en la forma de onda sea semejante, por lo que a priori son buenos candidatos para compartir modelado.

En concreto, el conjunto de cuestiones fonéticas que se utilizan en el presente proyecto fin de carrera, al igual que el agrupamiento inicial de partida, fue diseñado por Salvador Alcón

Paniagua, en su proyecto fin de carrera “Diseño de un reconocedor automático de habla espontánea en castellano”. Dichas cuestiones se realizaron teniendo en cuenta el conjunto de reglas fonéticas de la lengua española y la organización en distintas clases de fonemas que se tienen en ella.

Aunque el proceso de creación del árbol binario se ha explicado para el agrupamiento de trifonemas que comparten el mismo fonema central, en realidad el agrupamiento utilizado es un poco más complejo. En lugar de obtener un único árbol por cada grupo de trifonemas inicial, se obtienen 3 árboles diferentes, uno para cada estado emisor del HMM. De esta forma, se tendrán estados compartidos entre distintos HMMs. Cuando todos los estados de 2 o más HMMs sean compartidos, estos HMMs se consideran atados.

Con los datos de entrenamiento de la base de datos MICROAES, y utilizando un valor de 350 para el umbral que marca el aumento mínimo de la log-verosimilitudes y un valor de 100 para el umbral que marca el número de muestras mínimas para cada agrupación, se ha conseguido reducir el número de HMMs a modelar hasta los 8485, siendo el número de partida de 51987 (teniendo en cuenta monofonos, bifonemas (cuando sólo existe un contexto) y trifonemas).

Ya que las modificaciones aquí realizadas pueden generar cambios bastante significativos, tras obtener el árbol de trifonemas se reestiman los nuevos modelos mediante 4 ejecuciones del algoritmo de Baum-Welch.

Una segunda fase del entrenamiento de trifonemas consiste en aumentar el número de gausianas que modelan la probabilidad de emisión de cada estado emisor de los HMMs.

Este incremento se irá realizando incrementalmente, reestimándose los modelos mediante 4 ejecuciones del algoritmo de Baum-Welch en cada incremento.

En primer lugar se pasa de mezclas de 1 gausiana por estado, a mezclas de 2 gausianas, después se pasa de 2 a 4 gausianas por mezcla y después de 4 a 8 gausianas. Este proceso es común a todos los HMMs del sistema.

Por último, sólo para los HMMs encargados de modelar transiciones (modelo de basura, ‘sil’ y ‘sp’), se pasa de 8 a 16 gausianas, y de 16 a 32. Este aumento se realiza sólo en estos modelos especiales porque van a modelar eventos que pueden tener realizaciones muy diferentes, y podrían obtenerse mejores resultados si se le ofrece mayor capacidad de modelado.

d) Reconocimiento

Se considera que la última fase de un reconocedor es el propio reconocimiento. En este apartado se pretende explicar el proceso seguido por el reconocedor para reconocer una determinada locución a partir de los HMMs obtenidos en el entrenamiento. Además de ofrecer una salida con la locución reconocida, HTK permite obtener estadísticas de los resultados, cuya interpretación básica también será mencionada.

En el reconocimiento se parte de una secuencia de muestras de voz parametrizadas y se pretende obtener la trascripción de la locución que las generó.

Para indicar las posibles secuencias de palabras a reconocer se utiliza un modelo de lenguaje. Al no disponer de ningún modelo de lenguaje en castellano, se utiliza uno equiprobable, lo que indica que dada una palabra, la probabilidad de que se genere tras ella cualquiera de las palabras que componen el modelo de lenguaje es la misma. Ya que no se está aprovechando el gran potencial que pueden ofrecer los modelos de lenguaje a la hora de impedir transiciones extrañas, lo que si se ha hecho es limitar el conjunto de palabras que lo forman a justo las presentes en las frases de test, lo que impedirá que se reconozcan palabras que no se dan en dichas frases.

En el caso bajo estudio, ya que sólo existen 2 frases distintas de test (que repiten todos los locutores), el número de palabras que forman el modelo de lenguaje es de 145, un número bastante reducido que permitirá un reconocimiento bastante rápido.

En la Figura 17 se muestra de forma gráfica la forma del modelo de lenguaje utilizado, suponiendo que las locuciones de entrenamiento son ‘la casa’ y ‘la calle’. Como se puede comprobar, el modelo de lenguaje estará formado por las tres palabras que forman las frases de test, añadiendo de forma opcional al principio y final de frase el silencio largo ‘sil’, por si se produjeran pausas prolongadas en los límites de la grabación.

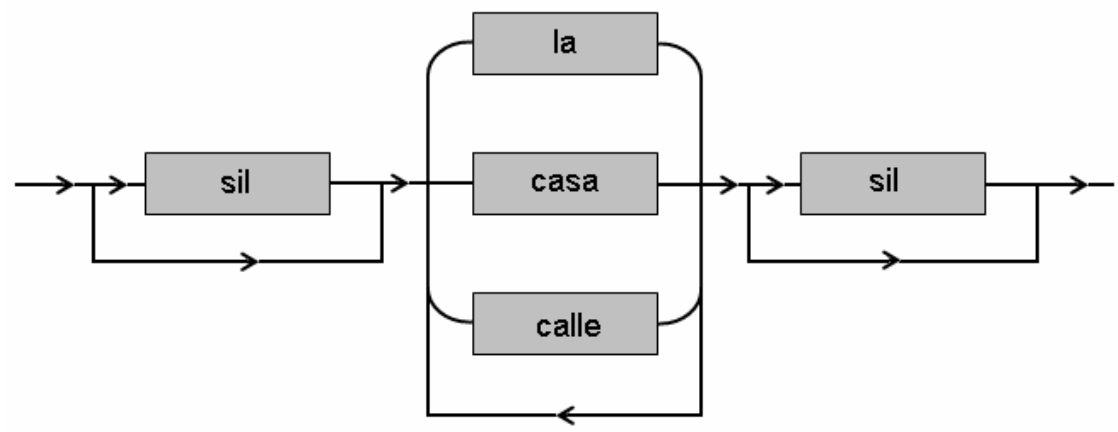

Figura 17.- Ejemplo de modelo de lenguaje equiprobable

Por lo tanto, el modelo de lenguaje indicará todos las posibles locuciones a reconocer, definida cada una de ellas como un camino seguido a lo largo del modelo de lenguaje. El reconocimiento consiste en obtener el camino con mayor probabilidad dada la locución parametrizada, obteniéndose así la secuencia de palabras que forman la trascripción.

Una vez se tienen todos los posibles caminos, con el diccionario se obtienen la transcripción de cada palabra. En el caso actual, cada palabra tiene tres posibles trascripciones, cada una terminada con un modelo de pausa diferente. Hay que evaluar todas las opciones, por lo que es como si se tuviera un modelo de lenguaje con el triple de palabras.

Una vez se tiene la transcripción de las palabras, hay que transformar cada posible camino del modelo de lenguaje en su correspondiente secuencia de trifonemas, que es la unidad acústica que se está modelando en los HMMs. Debido a que se está utilizando un árbol de trifonemas, para obtener los HMMs a partir de los trifonemas hay que tenerlo en cuenta para buscar el HMM (atado o no) correspondiente.

Finalizado este proceso, se tiene representado cada posible camino por una secuencia de HMMs, con sus probabilidades de emisión y de transición entre estados. Pasar de una palabra a la siguiente estará penalizado con un factor conocido como WIP (penalización por inserción de palabra). Además, cada palabra estará marcada según la probabilidad del modelo de lenguaje para esa palabra, que en este caso es la misma para todas las palabras, y un factor, LM , que indica la importancia que se le da a esta probabilidad.

El proceso de reconocimiento consistirá en calcular la probabilidad de la locución de entrada dado cada uno de los posibles caminos que puedan generarla (teniendo en cuenta que se debe pasar por un total de T estados emisores, siendo T el número de vectores de parámetros obtenidos a partir de la locución), sabiendo que cada palabra que atraviesa la locución influye en la probabilidad final a maximizar en un factor indicado en (3.1):

$$P_{w_i} = LM \cdot P_{LM}(w_i) + WIP + P(w_i) \qquad (3.1)$$

Siendo $P(w_i)$ la probabilidad acústica de la palabra w_i, cuyo valor dependerá de la probabilidad acústica de todos los HMMs que formen dicha palabra. La probabilidad acústica de un determinado HMM será la probabilidad de que la secuencia de observaciones que se ha decidido que pertenezcan a ese HMM haya sido generada por él, teniendo en cuenta las probabilidades de transición y de emisión de los distintos estados que lo forman.

Todo este proceso es desarrollado por la herramienta HVite de HTK, que teniendo en cuenta el conjunto de modelos, el diccionario y el modelo de lenguaje, utiliza una variación del algoritmo de Viterbi, conocido como algoritmo de paso de token (Token Passing) para obtener el camino con mayor probabilidad, y con esto, la trascripción de la locución bajo test.

Por último comentar que, sólo para los casos en los que están disponibles las transcripciones reales de las locuciones de test, HTK proporciona una herramienta (HResults) para evaluar los resultados.

Esta herramienta compara el conjunto de transcripciones resultado del reconocimiento con las transcripciones de referencia, utilizando para ello un algoritmo basado en programación dinámica. Con esto, proporciona estadísticas sobre resultados a nivel de locución y a nivel de palabra, utilizando el formato mostrado en la Figura 18:

```
---------- Overall Results ----------
SENT:   %Correct=13.00          [H=13, S=87, N=100]
WORD:   %Corr=53.36, Acc=44.90  [H=460, D=49, S=353, I=73, N=832]
==========================================
```

Figura 18.- Ejemplo de salida de HResult

En cuanto a lo información proporcionada a nivel de locución (línea *SENT* de la Figura 18) el parámetro *%Correct* indica, en tanto por ciento, las locuciones que son idénticas a sus locuciones de referencia, *H* es el número de locuciones correctas, *S* el número de locuciones con errores y *N* el número total de locuciones. La expresión que permite obtener el valor de %Correct sería la indicada en (3.2):

$$\%Correct = \frac{H}{N} \times 100 \qquad (3.2)$$

En cuanto a lo información proporcionada a nivel de palabra (línea *WORD* de la Figura 18) el parámetro *%Corr* indica, en tanto por ciento, las palabras que coinciden exactamente con las de las locuciones de referencia, *Acc* indica lo mismo que el parámetro anterior, pero teniendo en cuenta el número de inserciones, *H* es el número de palabras correctas, *D* el número de palabras eliminadas, *S* el número de sustituciones, *I* el número de inserciones y *N* el número total de palabras de las transcripciones reales. La expresión que permite obtener el valor de %Corr coincide con la Fórmula 3.2, y la expresión para el cálculo de Acc se indica en (3.3).

$$Acc = \frac{H - I}{N} \times 100 \qquad (3.3)$$

Además, esta herramienta permite ignorar ciertas etiquetas, para no tenerlas en cuenta en la evaluación de los resultados. En el presente reconocedor, se ignora la etiqueta correspondiente al modelo basura al analizar los resultados, puesto que se considera que no aporta información el hecho de que se haya reconocido correctamente o no. También se ignora la etiqueta del modelo 'sil' del principio y final de la locución, que puede aparecer o no en el reconocimiento, y que tampoco aportará información.

3.3.6 Experimentos realizados

Sobre el reconocedor desarrollado se ha realizado un único experimento para obtener los resultados de reconocimiento que ofrece sobre el conjunto de datos de test de la base de datos utilizada (MICROAES).

Este experimento será el punto de partida con el que poder comparar futuros resultados de reconocimiento, fruto de diferentes modificaciones que se realizarán sobre la implementación básica del reconocedor, con el fin de introducir información sobre los locutores en el proceso y observar de qué forma influye esta información en los resultados.

A continuación se resumen los parámetros fijados sobre el reconocedor para esta prueba concreta:

- Tanto los datos de entrenamiento como de test son los que indica la base de datos. En concreto se tendrán 300 locutores que aparecen en ambos grupos de datos, teniendo 15 frases de entrenamiento y 2 de test por cada uno de ellos, grabadas por 2 micrófonos de forma simultánea. Por lo tanto, se tendrán 30 locuciones de entrenamiento y 4 de test para 300 locutores diferentes.
- Se utiliza el diccionario aportado por la base de datos, con las modificaciones ya mencionadas en 3.3.2.
- Se consideran un total de 36 fonemas diferentes, que se indican en la Tabla 5.
- Además de los modelos correspondientes a los fonemas, se tienen 2 modelos para modelar silencios (modelo 'sil' para silencio largo y modelo 'sp' para silencio corto) más un tercer modelo para las dudas, palabras incompletas o ruidos del locutor.
- El resultado de la fase de entrenamiento será un total de 8485 HMMs, conjunto donde se encuentran los 3 modelos especiales utilizados para la separación entre palabras. El resto es un conjunto de *trifonemas tree-clustered tied-state* que, gracias al árbol de agrupación, sirve para modelar todos los posibles trifonemas que se puedan originar.

- En concreto, el árbol de agrupamiento se obtuvo gracias a las cuestiones fonéticas diseñadas por Salvador Alcón Paniagua, utilizando un valor de 350 para el umbral que marca el aumento mínimo de la log-verosimilitud y un valor de 100 para el umbral que marca el número de muestras mínimas para cada agrupación.
- Para el reconocimiento se utiliza un modelo de lenguaje equiprobable, formado por 145 palabras más el modelo de silencio largo 'sil' opcional al principio y final.
- La importancia del modelo de lenguaje se fija mediante un peso de 16.0, y la penalización por inserción de palabra será de -55.
- En los resultados de reconocimiento no se tendrá en cuenta la aparición de las etiquetas correspondientes a los modelos 'sil' o basura (etiqueta 'hesitation') al comparar las transcripciones correspondientes al reconocimiento y las reales.

En este experimento de referencia el reconocimiento se realiza con el conjunto de modelos que utilizan mezclas de 8 gausianas por estado para los modelos de trifonemas y de 16 para los modelos de silencio y el modelo basura.

De todos los valores que ofrece la herramienta HResults de HTK, para evaluar los resultados sólo se utilizará el parámetro *Acc*, ya que da una medida más global sobre la exactitud de reconocimiento, al tener en cuenta también el número de inserciones erróneas que se introducen en las transcripciones.

El resultado obtenido para esta primera prueba, y que deberá tenerse presente durante el resto del proyecto, es de:

Acc=73.12% [H=54258, D=5260, S=11394, I=2410, N=70912]

Resultado1.- prueba de referencia

Aunque el reconocedor está preparado para entrenar el modelo basura con 32 gausianas, finalmente se ha usado un modelo de 16. Se ha elegido así porque en el proyecto de Salvador Alcón Paniagua se hizo una batería de pruebas sobre el reconocedor, obteniéndose la conclusión de que las mejoras sobre el reconocimiento al aumentar de 16 a 32 el número de gausianas del modelo basura eran poco significativas. Este efecto posiblemente se produzca porque se realiza un sobreentrenamiento del modelo sobre los pocos datos de entrenamiento existentes. Como no produce mejoras, se utiliza el modelo de 16 gausianas, que reduce el tiempo de entrenamiento respecto al uso del modelo más complejo.

Capítulo 4

Efectos del género del locutor sobre el reconocimiento de habla

4.1 Resumen

En este capítulo se pretende observar en qué medida se ven afectados los resultados de reconocimiento de habla si se utiliza información sobre el género de los locutores tanto en la fase de entrenamiento como en la de test.

Para que se trate de un sistema con aplicación real se va a suponer que el género de las locuciones de test no es conocido, por lo que será necesario el desarrollo de un sistema que permita obtener el género del locutor que las generó.

Una vez se dispone de las herramientas necesarias para diferenciar el género de las locuciones, ya sea a través de la información real obtenida de la base de datos o bien a través

del clasificador de género, se plantea el problema de introducir esa información en el reconocedor.

Hasta ahora el reconocedor de habla trabaja con un único conjunto de modelos, generado a partir de la totalidad de las locuciones de entrenamiento. En este momento se plantea la obtención de dos conjuntos diferentes, donde cada uno de ellos representará la señal de voz para un determinado género.

La obtención de este conjunto de modelos se plantea de dos formas diferentes:

- Realizando el mismo entrenamiento ya definido, pero para cada conjunto de modelos, utilizando bien los datos de los locutores masculinos o bien los de los femeninos. Para que los modelos representen correctamente a cada grupo, los datos de entrenamiento deben ser suficientemente extensos. Con la base de datos MICROAES, al dividir los datos de entrenamiento en función del género, se está dividiendo aproximadamente a la mitad la cantidad de información utilizada para cada grupo respecto a la disponible para el experimento de referencia.
- Debido a la limitación en el número de datos de entrenamiento, se plantea el uso de adaptación sobre el conjunto de modelos ya obtenido, que estará suficientemente bien entrenado puesto que se ha usado el conjunto completo de datos para su generación. La adaptación MAP (Bayesian Maximum A Posteriori) será la técnica elegida, puesto que a priori ofrece buenos resultados cuando la cantidad de datos de adaptación es alta, como es el caso actual (aproximadamente la mitad de los datos de entrenamiento de la base de datos para cada conjunto).

Tras una introducción teórica sobre la técnica de adaptación utilizada, se explicarán los desarrollos llevados a cabo sobre el clasificador de género, seguido por la batería de pruebas realizadas sobre el reconocedor, indicándose a posteriori las conclusiones obtenidas.

4.2 Adaptación MAP

Como ya se ha comentado, se pretende conseguir dos grupos de modelos, uno que caracterice a los locutores masculinos y otro que caracterice a los femeninos. Entrenar estos modelos desde el principio, utilizando los datos separados por género, presenta un gran inconveniente: la gran cantidad de datos necesaria para el entrenamiento.

Para solucionar este inconveniente se utiliza la adaptación [YEG+06], [HAH01]. Partiendo de un grupo de modelos inicial bien entrenado, se generan los nuevos modelos adaptando los primeros mediante los datos correspondientes a cada uno de los grupos. La cantidad de datos necesaria en este proceso es menor que la que se necesita para entrenar los modelos desde el principio, puesto que ya se conoce en torno a qué valores se van a encontrar los parámetros de los modelos, y se trata únicamente de ajustarlos a cada grupo.

Tanto la cantidad de datos necesaria como los resultados del reconocimiento realizado con estos modelos adaptados dependerán del tipo de adaptación elegida.

La adaptación MAP es una de las técnicas usada en adaptación de HMMs, y será la utilizada en este proyecto para realizar la adaptación al género, eligiéndose su versión supervisada, lo que implica que se conoce la trascripción correcta de los datos que intervienen en el proceso. Cuando esta información es desconocida se hace necesario el uso de adaptación no supervisada, donde se estima de alguna manera esta información no presente.

La adaptación MAP [GL94] se basa en emplear el conocimiento a priori de los modelos, ajustando sus parámetros a las características de los nuevos datos. Este conocimiento a priori sobre el modelo que se está adaptando impide que se produzcan grandes desviaciones de sus parámetros, que se darán únicamente si los nuevos datos de adaptación proporcionan fuertes indicios de ello.

En un entrenamiento normal se obtienen los modelos de forma que se maximice $P(O|M)$, siendo O una determinada secuencia de observaciones producidas por el modelo M que se está entrenando. Con la adaptación MAP lo que se pretende es introducir una probabilidad a priori de dicho modelo, para obtener una estimación más exacta de los parámetros del modelo.

$$M_{MAP} = \arg\max_{M} P(M|O) = \arg\max_{M} P(O|M)P(M) \qquad (4.1)$$

Siendo O las datos de adaptación del modelo M .

En el caso de que no exista información a priori para alguno de los modelos a adaptar, se considera esa información como una constante, con lo que no interviene en la maximización de (4.1), y por lo tanto la estimación MAP obtenida coincidirá con una estimación de máxima verosimilitud, es decir, el valor que se obtendría si se estuviera realizando un entrenamiento normal con esos pocos datos de adaptación.

La fórmula de actualización [YEG+06], [HAH01] de la media de la gausiana *m* del estado *j* de un determinado HMM es la que se indica en (4.2):

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm} + \tau} \overline{\mu}_{jm} + \frac{\tau}{N_{jm} + \tau} \mu_{jm} \qquad (4.2)$$

Siendo $\overline{\mu}_{jm}$ el valor de la media para los datos de adaptación, μ_{jm} el valor de la media del modelo a priori y $\hat{\mu}_{jm}$ el valor de la media tras la adaptación MAP. N_{jm} y τ son los encargados de dar mayor o menor peso a los valores de $\overline{\mu}_{jm}$ y de μ_{jm}. N_{jm} es la probabilidad de ocupación de los datos de adaptación, por lo que no es un valor configurable, calculándose según (4.3):

$$N_{jm} = \sum_{t=1}^{T} L_{jm}(t) \qquad (4.3)$$

Siendo $L_{jm}(t)$ la probabilidad de encontrarse en la mezcla *m* del estado *j* del modelo en el momento *t*. El valor de la media de los datos de adaptación se calcula en función de este parámetro:

$$\overline{\mu}_{jm} = \frac{\sum_{t=1}^{T} L_{jm}(t) o_t}{\sum_{t=1}^{T} L_{jm}(t)} \qquad (4.4)$$

En cuanto a τ, es un parámetro fijado por configuración, permitiendo aumentar o disminuir la importancia del conocimiento a priori para un valor de N_{jm} dado. Si el valor de τ fijado es mucho mayor que N_{jm}, según expresa (4.2) la adaptación estará gobernada por la información a priori, mientras que si τ tiene un valor muy inferior a N_{jm}, la media adaptada será muy próxima al valor de la media para los datos de adaptación, sin tener en cuanta la información a priori existente.

La matriz de covarianza resultante de la adaptación MAP se obtiene a partir del valor de la media adaptada, como se indica en (4.5):

$$\hat{\Sigma}_{jm} = \frac{S_{jm} + \tau(\hat{\mu}_{jm} - \mu_{jm})(\hat{\mu}_{jm} - \mu_{jm})' + \sum_{t=1}^{T} N_{jm}(o_t - \hat{\mu}_{jm})(o_t - \hat{\mu}_{jm})'}{\eta_{jm} - D + N_{jm}} \qquad (4.5)$$

Siendo S_{jm} un parámetro dependiente de la matriz de covarianza del modelo a priori, η_{jm} un parámetro relacionado con τ, y por lo tanto, relacionado con la importancia que se le da a la información a priori, y D la dimensión del vector de medias.

Aunque no se ha especificado la formulación matemática de la adaptación del resto de parámetros de un HMM (pesos de las gausianas que componen cada estado y matrices de transición), mencionar que se puede realizar dicha adaptación, aunque no se va a utilizar en el presente proyecto, motivo por el que no se muestran las fórmulas correspondientes [GL94]. En cuanto a los pesos de las gausianas, indicar también que HTK no permite su adaptación MAP.

Tanto en (4.2) como en (4.5) se puede observar que el peso de la información a priori, τ, no depende ni de la gausiana ni del estado del modelo. En realidad se puede fijar este parámetro para cada gausiana a adaptar, pero se ha decidido mostrar uno único porque es así como se utilizará en el reconocedor del proyecto, es decir, en todas las adaptaciones el peso de la información a priori será el mismo.

4.3 Clasificador de género

El clasificador de género desarrollado consiste en un mecanismo automático para decidir si una determinada locución de entrada ha sido realizada por un locutor masculino o femenino.

Aunque la información de género está disponible para todas las locuciones de la base de datos MICROAES, con el objetivo de hacer más general el reconocedor de habla que se está diseñando, de forma que se pueda utilizar en otros entornos donde no se tenga esta información sobre el género, se decidió implementar un clasificador de género. En la fase de reconocimiento, este clasificador permitirá obtener el género de la locución de entrada, lo que permitirá elegir el conjunto de modelos adecuado para realizar la decodificación acústica.

Aunque éste fue el punto de partida que llevó al desarrollo del clasificador, finalmente se ha usado también en la fase de entrenamiento para distintas pruebas planteadas, de forma que se pueda deducir si su uso en dicha fase es favorable o perjudicial.

4.3.1 Principios de desarrollo

Para la realización del clasificador de género se ha partido de un software de verificación de locutor desarrollado por *Darío Martín Iglesias*, dentro del Grupo de Procesado Multimedia del Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid [Mar06], que se ha podido adaptar con bastante facilidad gracias a que los principios que utiliza son parecidos a los necesarios para el clasificador.

La tarea de verificación de locutor [RQD00] implica decidir si el locutor hipotético es realmente quien ha generado la frase de test, o si, por el contrario, no lo ha hecho. Matemáticamente se podría expresar como indica (4.6):

$$\frac{p(Y|H_0)}{p(Y|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \theta \qquad (4.6)$$

Siendo Y la locución bajo test, H_0 la hipótesis que indica que Y ha sido generado por el locutor hipotético, H_1 la hipótesis de que no ha sido generado por dicho locutor y θ el umbral de aceptación. Por lo tanto, si la probabilidad de que se haya generado Y dado que se tiene la hipótesis H_0 es mayor que el umbral θ por la probabilidad de que se haya generado Y dado que se tiene la hipótesis H_1, se elegirá como cierta la hipótesis H_0, y por lo tanto se verifica el locutor. En caso contrario se decide H_1, que indica que el locutor no es quien dice ser.

El primer tratamiento que se realiza es obtener la parametrización de la señal de voz, de forma que se extraen las características de dicha señal, que contendrán información sobre la identidad del locutor. Una vez realizada, la hipótesis H_0 se representará por un modelo matemático que caracterice al locutor hipotético en el espacio del vector de parámetros que caracteriza la señal de entrada, y H_1 vendrá representado por otro modelo de las mismas características pero que represente la opción alternativa al locutor hipotético. Por lo tanto, en la verificación, lo que se compara es la probabilidad de que la secuencia de observaciones parametrizadas haya sido generada por uno u otro modelo, eligiéndose el adecuado en función del umbral.

Esta forma de verificación de locutor es una particularización, para el caso de tener sólo dos modelos, de la fórmula utilizada para identificación de locutor [RR95], donde se debe elegir el locutor, dentro de un conjunto dado, que generó la voz de entrada (4.7).

$$locutor\_elegido = \arg\max_{1 \le S \le K} P(\lambda_{loc\_s}|O) = \arg\max_{1 \le S \le K} \frac{P(O|\lambda_{loc\_s}) \cdot P(\lambda_{loc\_s})}{P(O)} \qquad (4.7)$$

Siendo S el grupo de locutores disponibles, K la dimensión de S, O la señal de voz bajo test parametrizada, y λ_{loc_s} el modelo correspondiente al locutor 'loc_s'.

En (4.7) se indica que se obtendrá el locutor para el que se maximice la probabilidad del modelo que representa a dicho locutor dado que se tiene la secuencia de vectores de observación O. Utilizando la regla de Bayes se obtiene la expresión que permite obtener dicha probabilidad en función de la probabilidad de O dado el modelo del locutor, la probabilidad a priori de dicho modelo y la probabilidad de la secuencia de observaciones. Este último factor, $P(O)$, permanece constante para todos los locutores, por lo que no afectará en la elección del máximo. Además, si se tienen locutores equiprobables, la expresión se reduce a (4.8):

$$locutor\_elegido = \arg\max_{1 \le S \le K} P\left(O \middle| \lambda_{loc\_s}\right) \qquad (4.8)$$

En cuanto al problema de clasificación de género, se puede considerar como una particularización del de identificación de locutor. Se tendrán dos modelos, λ_{masc} que representará las características de los locutores masculinos, y λ_{fem} que representará la de los femeninos. Dada una locución de entrada, después de parametrizada y suponiendo que las probabilidades a priori de ambos modelos coinciden, se decidirá el modelo que con mayor probabilidad haya generado dicha entrada, proceso que queda representado en (4.9):

$$P\left(O \middle| \lambda_{masc}\right) \underset{fem}{\overset{masc}{\gtrless}} P\left(O \middle| \lambda_{fem}\right) \qquad (4.9)$$

En el caso de que las prioridades a priori de los modelos no sean iguales, se decidirá en función de un umbral. El siguiente desarrollo muestra el valor teórico del umbral teniendo en cuenta que se utilizan logaritmos.

$$P(\lambda_{masc}) P\left(O \middle| \lambda_{masc}\right) \underset{fem}{\overset{masc}{\gtrless}} P(\lambda_{fem}) P\left(O \middle| \lambda_{fem}\right) \qquad (4.10)$$

$$\ln\left(P(\lambda_{masc})\right) + \ln\left(P\left(O \middle| \lambda_{masc}\right)\right) \underset{fem}{\overset{masc}{\gtrless}} \ln\left(P(\lambda_{fem})\right) + \ln\left(P\left(O \middle| \lambda_{fem}\right)\right) \qquad (4.11)$$

$$\ln\left(P\left(O \middle| \lambda_{masc}\right)\right) - \ln\left(P\left(O \middle| \lambda_{fem}\right)\right) \underset{fem}{\overset{masc}{\gtrless}} \theta \qquad (4.12)$$

$$\theta = \ln\left(P\left(\lambda_{fem}\right)\right) + \ln\left(P\left(\lambda_{masc}\right)\right) \qquad (4.13)$$

Por lo tanto, queda probado que la clasificación de género es un caso especial de clasificador de locutor, quedando el género representado con dos modelos diferentes. Puesto que el problema de verificador de locutor es también un caso especial de identificación de locutor, utilizando únicamente dos modelos, si se utiliza el verificador de locutor de partida con modelos entrenados en función del género de los locutores se obtiene el clasificador de género buscado.

El software de partida es un verificador de locutor independiente del texto, modelando la identidad de cada locutor mediante GMMs (*Gaussian Mixture Models*). En primer lugar se entrena un modelo general, conocido como modelo 'mundo', utilizando un conjunto adecuado de los datos de entrenamiento, para después adaptar este modelo para cada uno de los locutores mediante adaptación MAP, utilizando los datos de adaptación de ese locutor. Cada modelo se corresponde con un GMM, y el número de gausianas que lo compone vendrá marcado por configuración.

Una vez realizado el entrenamiento, la verificación del locutor bajo test se acepta si se cumple la siguiente indicada por (4.14):

$$\ln\left(P\left(O\middle|\lambda_{locutor}\right)\right) - \ln\left(P\left(O\middle|\lambda_{mundo}\right)\right) \underset{mundo}{\overset{locutor}{\rangle}} \theta \qquad (4.14)$$

Siendo O la señal de voz parametrizada, $\lambda_{locutor}$ el GMM que modela al locutor hipotético y λ_{mundo} el modelo universal alternativo.

Si la anterior expresión no supera el umbral, θ, el usuario es rechazado. El valor del umbral debería ser elegido en función de las probabilidades a priori de los usuarios.

La adaptación de este software para el caso de clasificación de género consiste en entrenar el modelo 'mundo' con todos los datos de entrenamiento, realizando una adaptación MAP del mismo con el conjunto de entrenamiento de los locutores masculinos y otra con el de los femeninos, obteniendo así dos modelos que identificarán el género del locutor. A continuación, para cada locución de test, se calculará la probabilidad de que dicha locución se haya generado por uno y otro modelo, decidiendo el género del locutor en función del umbral que se indique para comparar dichas probabilidades.

4.3.2 Estructura del clasificador de género

Igual que en el reconocedor, para el desarrollo del clasificador de género se utilizó lenguaje de scripting bajo Linux para manejar las herramientas ofrecidas por HTK.

La base de datos que se va a utilizar será la misma que en el reconocedor (MICROAES). Se van a utilizar dos divisiones de dicha base de datos para realizar distintas pruebas al clasificador.

En la primera división de los datos se eligen dos grupos con locutores independientes, es decir, los locutores que forman los datos de test no aparecen como locutores del conjunto de entrenamiento.

La segunda división de los datos es la ofrecida por la propia base de datos, y la que se ha usado también en el reconocedor.

Debido a que el objetivo principal del desarrollo del clasificador de género es poder obtener, en la fase de test del reconocedor, el género al que pertenece la locución que se va a reconocer, aunque se intente conseguir unos modelos de género entrenados con datos independientes del locutor, esto carece de sentido al utilizarlo en reconocimiento, ya que los 300 locutores disponibles en la base de datos están representados en las locuciones de test del reconocedor.

Por lo que se acaba de mencionar, el clasificador de género que se va a utilizar finalmente junto con el reconocedor será el diseñado a partir de la división en datos de entrenamiento y test que recomienda MICROAES. A pesar de esto, se ha realizado una batería de pruebas al clasificador con la división de datos independientes del locutor, con el fin de observar si esta característica influye en los resultados.

El primer paso del clasificador de género, una vez elegida la base de datos, es realizar la parametrización de los archivos de voz que van a intervenir en el proceso. Con esto se pretende obtener las características más significativas de la señal de voz, además de reducir espacio de almacenamiento. En concreto, todo el proceso se basa en que en dichas características esté representado el género de los locutores. Posteriormente, se comprobará si diferenciar la generación de los modelos del reconocedor en función del género de los locutores aporta o no beneficios en el proceso de reconocimiento, lo cual indicaría que la suposición inicial es cierta y, por lo tanto, las características de la señal de voz tienen cierta dependencia del género.

En concreto, la parametrización utilizada para el clasificador de género es la misma que la del reconocedor de voz.

En cuanto a los modelos utilizados, ya se ha comentado que se tendrán tres modelos diferentes:

- Un modelo universal o modelo 'mundo' (λ_{mundo}), que será entrenado con el total de locuciones de entrenamiento.
- Un modelo masculino (λ_{masc}), que será una adaptación del modelo 'mundo' a los datos de entrenamiento de locutores masculinos.
- Un modelo femenino (λ_{fem}), que será una adaptación del modelo 'mundo' pero esta vez para los locutores femeninos.

Se han elegido los GMMs como modelos a utilizar. Son modelos constituidos por sumas de gausianas (mezclas de gausianas), por lo que se trata de una particularización de los HMMs para el caso de tener un único estado emisor. La topología utilizada queda representa en la Figura 19.

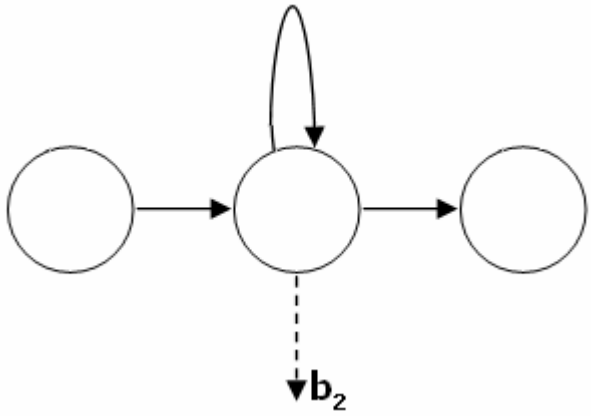

Figura 19.- Topología de los GMMs del clasificador de género

Una vez elegida la forma de los modelos y los datos que representan a cada uno de ellos, con sus correspondientes transcripciones indicando el género al que pertenecen, comienza la fase de entrenamiento.

Esta fase consta de dos tareas:

- Entrenamiento del modelo 'mundo'. En este entrenamiento será donde se defina el número de gausianas que van a componer los GMMs obtenidos en todo el proceso.
- Adaptación del modelo anterior al género, utilizando para ello adaptación MAP. Tanto para λ_{masc} como para λ_{fem} se generará un GMM con el mismo número de gausianas que el modelo 'mundo'.

La última fase es el reconocimiento. Éste se realiza indicando el umbral (θ) a utilizar en la comparación de las verosimilitudes. Para cada archivo de test se calculará la probabilidad de que

haya sido generado tanto por el modelo masculino (λ_{masc}) como por el femenino (λ_{fem}), indicándose el modelo elegido tras la decisión. Durante la ejecución se imprimen en pantalla las verosimilitudes obtenidas y las estadísticas de aciertos.

La Figura 20 muestra un esquema de las fases del clasificador de género.

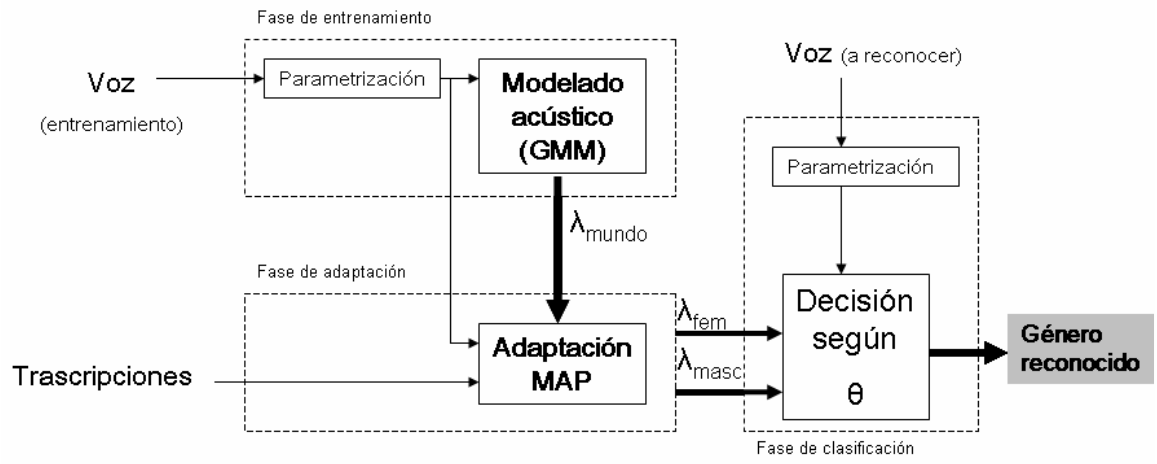

Figura 20.- Clasificador de género

Para poder obtener resultados sobre las prestaciones del sistema es necesario conocer el género real del locutor de cada una de las locuciones de test, por lo que será necesario generar las transcripciones de estos archivos, almacenando únicamente la etiqueta que indique el género.

4.3.3 Experimentos de clasificación de género

Los experimentos realizados sobre el clasificador de género se separan en dos grupos en función de la división realizada sobre la base de datos para obtener el conjunto de entrenamiento y de test.

Para comprobar el funcionamiento del clasificador de género se obtienen medidas sobre la probabilidad de acierto, utilizando la expresión (4.15) para calcularla:

$$\%Corr = \frac{H}{N} \times 100 \qquad (4.15)$$

Siendo H el número de locuciones reconocidas correctamente y N el número total de locuciones evaluadas.

Esta fórmula se aplicará en tres contextos diferentes:

- La probabilidad de acierto del sistema completo, para la que H se calcula como el número de locuciones para las que se ha obtenido el género correcto, y N es el total de locuciones de test.
- La probabilidad de acierto para los locutores masculinos, para la que H se corresponde con el número de locuciones masculinas para las que se ha obtenido dicho género, y N es el número total de locuciones masculinas de test.
- La probabilidad de acierto para los locutores femeninos, para la que H se corresponde con el número de locuciones femeninas para las que se ha obtenido dicho género, y N es el número total de locuciones femeninas de test.

Lo que se buscará será mejorar estas dos últimas probabilidades, de tal forma que sean lo más parecidas posibles, es decir, que se tenga la misma probabilidad de acertar un locutor masculino que uno femenino. El umbral será fijado para conseguir este objetivo, y no en función de la probabilidad a priori de los modelos.

4.3.3.1 Datos independientes del locutor

En este primer conjunto de experimentos la división en datos de entrenamiento y test utilizada no es la que se indica en las especificaciones de la base de datos. Se decidió utilizar una división que permitiera que los resultados no fueran dependientes del locutor. Para ello se utilizan los primeros 250 locutores de la base de datos para entrenamiento, y los 50 restantes para test, estando ambos grupos balanceados en género. Cada locutor aporta un total de 34 locuciones al grupo al que pertenece, tratándose de las 15 locuciones marcadas por la base de datos como ‘s’ (entrenamiento) más las 2 marcadas como ‘x’ (test), de los 2 micrófonos más cercanos al locutor.

Con los datos divididos de la forma anterior, se está asegurando que los locutores con los que se va a testear el sistema no han sido utilizados previamente en el entrenamiento, de forma que no se produzca también una adaptación a los locutores.

En primer lugar se hace una batería de pruebas para observar el efecto del número de gausianas de los GMMs, número que irá aumentando como potencia de 2, ya que esto es lo permitido por el clasificador de género. En todos estos experimentos el umbral utilizado será ‘0’, de forma que se elige para cada locución el modelo, de los dos posibles, con mayor probabilidad de generarla.

Una vez elegido el número de gausianas para las que se obtienen mejores prestaciones, se pasa a realizar un ajuste en el umbral para igualar la tasa de acierto en los dos tipos de locutores, lo que proporciona igualdad en la probabilidad de reconocer ambos géneros.

a) Ajuste al número de gausianas

Esta batería de pruebas se realiza utilizando varias versiones del entrenamiento del modelo 'mundo', a partir del que se obtienen λ_{masc} y λ_{fem}. Cada versión se caracteriza por el número de veces que se utiliza el algoritmo de Baum-Welch para hacer la reestimación del modelo 'mundo' en cada incremento del número de gausianas.

Una reestimación cada incremento de gausianas en el modelo 'mundo':

En este caso, en cada incremento del número de gausianas que forman el modelo, simplemente se ejecuta el algoritmo de reestimación en una ocasión. Los resultados se representan en la Figura 21 (datos en Tabla 10 del Anexo I).

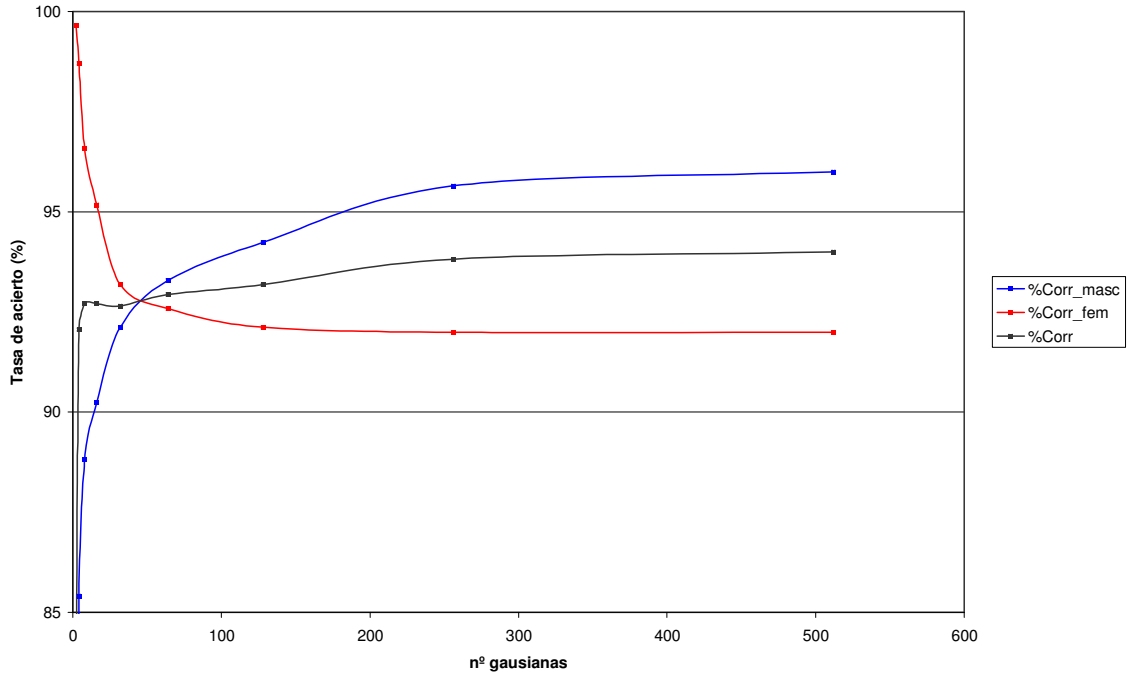

Figura 21.- Grupos independientes del locutor – 1 reestimación

En la Figura 21 se puede ver como un buen resultado sería elegir los modelos de 512 gausianas, puesto que aumenta ligeramente la tasa de aciertos de los locutores masculinos manteniéndose constante la de los femeninos. A pesar de esto, también se podría elegir el caso de 256 gausianas, puesto que se obtienen resultados parecidos y la complejidad tanto en el entrenamiento como en el test disminuye respecto a utilizar 512 gausianas.

Dos reestimaciones cada incremento de gausianas en el modelo 'mundo':

En este caso, en cada incremento del número de gausianas que forma el modelo, se ejecuta el algoritmo de reestimación en dos ocasiones. Los resultados se representan en Figura 22 (datos en Tabla 11 del Anexo I).

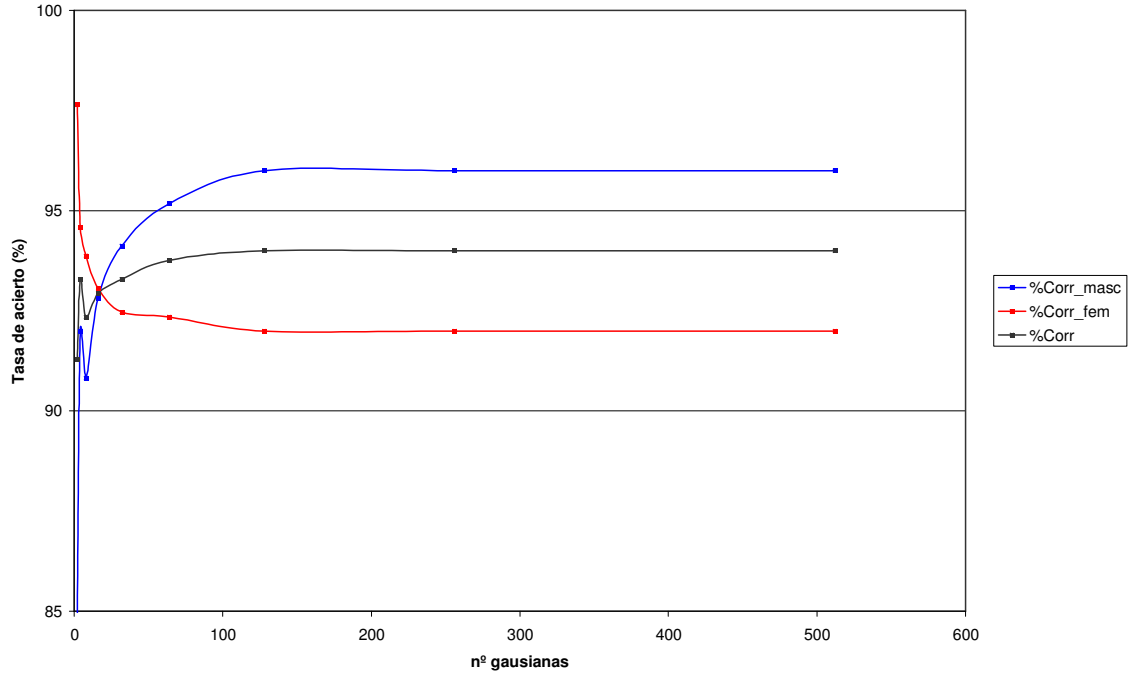

Figura 22.- Grupos independientes del locutor – 2 reestimaciones

Se puede comprobar que se alcanzan los resultados del caso anterior con un menor número de gausianas, es decir, las tasas de acierto que se obtenían con 512 gausianas para una reestimación se obtienen con 128 gausianas es el caso de realizar 2 reestimaciones. Por lo tanto, parece muy beneficioso aumentar a dos el número de ejecuciones del algoritmo de reestimación, ya que aunque aumenta el tiempo de entrenamiento necesario en cada incremento del número de gausianas, se pueden utilizar modelos más pequeños, que será siempre más eficiente para el sistema de reconocimiento global.

Como la mejora es significativa, se considerada adecuado realizar pruebas ejecutando en tres ocasiones el algoritmo Baum-Welch para reestimar.

Tres reestimaciones cada incremento de gausianas en el modelo 'mundo':

Al incrementar el número de gausianas se entrena el modelo tres veces. Los resultados se representan en la Figura 23 (datos en Tabla 12 del Anexo I).

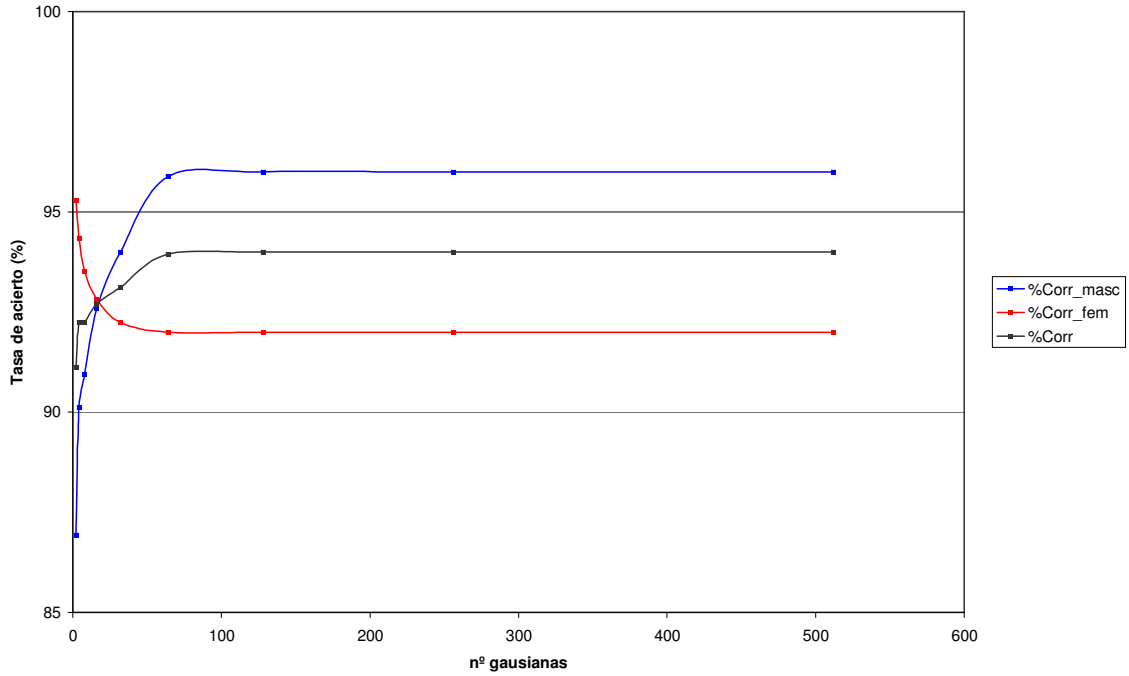

Figura 23.- Grupos independientes del locutor – 3 reestimaciones

A primera vista la mejora no es tan significativa como al pasar de una a dos reestimaciones, ya que hasta modelos de 128 gausianas no se obtienen los mismos resultados que en el caso anterior. Si se decide utilizar finalmente un modelo de 128 gausianas, obviamente debería elegirse el entrenamiento con dos reestimaciones, puesto que el coste de entrenamiento es menor. Sin embargo, para los modelos de 64 gausianas se puede observar que la tasa de aciertos femeninos es la misma que en el caso de 128 gausianas, y que el empeoramiento en los masculinos es muy pequeño (se pasa de un 96.00% a un 95.88% de aciertos). Por lo tanto, como los resultados son muy parecidos, habría que elegir entre estos dos modelos. Considerando como factor más importante la disminución de la complejidad de la fase de test, el utilizar modelos de 64 gausianas es más ventajoso que el uso de modelos de 128, aunque estos permitan disminuir el tiempo de la fase de entrenamiento utilizando sólo 2 reestimaciones por incremento.

Como el clasificador de género desarrollado se va a utilizar dentro de un reconocedor de habla, se considera que en la fase de test del reconocedor el tiempo que se tarda en decidir si el locutor es masculino o femenino debe ser lo menor posible, por lo que finalmente se elige utilizar un entrenamiento con reestimaciones con 3 ciclos del algoritmo de Baum-Welch, fijando el número de gausianas de los GMMs en 64.

b) Ajuste del umbral

Una vez elegido el número de gausianas de los modelos, debido a que la tasa de aciertos masculinos es mayor que la de femeninos, se procede a ajustar el umbral para que ambas se compensen. Los resultados se representan en la Figura 24 (datos en Tabla 13 del Anexo I).

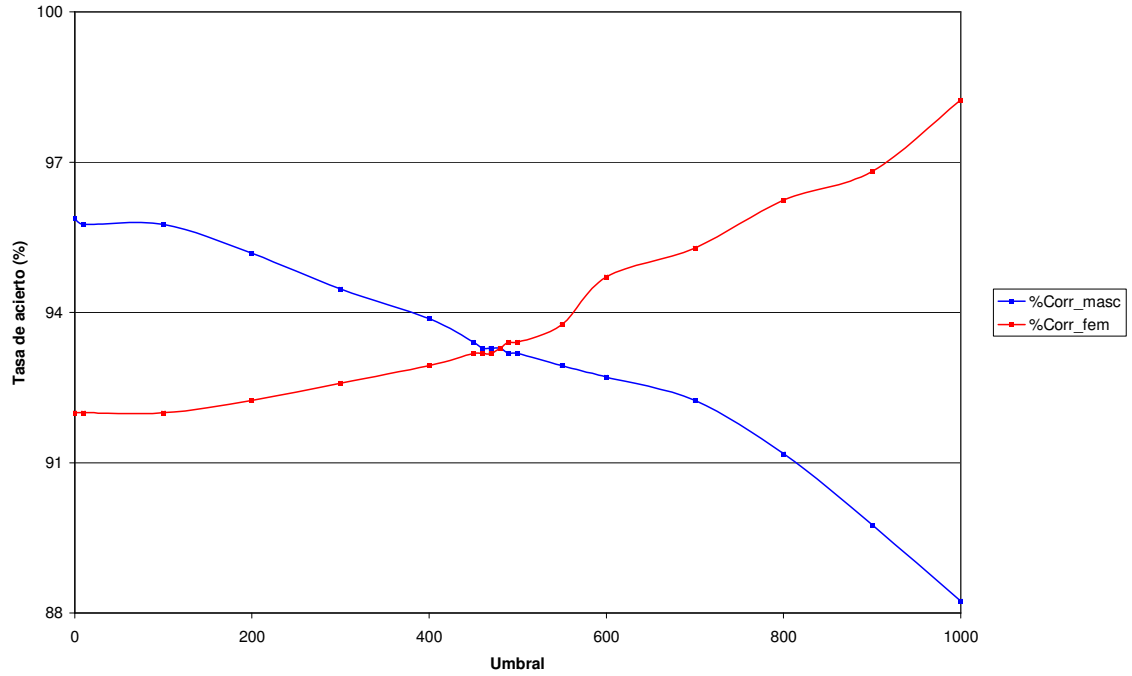

Figura 24.- Grupos independientes del locutor – ajuste del umbral

Aunque el barrido de valores fue superior, en la Figura 24 sólo se muestra el rango de 0 a 1000. La representación se fija de esta forma porque para elegir el umbral se pretende igualar la tasa de aciertos masculinos y la de femeninos, y es en este rango del umbral donde se toman mayores valores de ambas medidas conjuntamente.

El umbral óptimo obtenido es de 480, valor para el que la tasa de aciertos masculinos es igual a la de femeninos, siendo ambas de 93.29%.

4.3.3.2 Datos dependientes del locutor

En este segundo conjunto de experimentos la división en datos de entrenamiento y test utilizada es la que se indica en la base de datos, es decir, tanto en el entrenamiento como en el test estarán presentes los 300 locutores, aportando cada uno 30 locuciones de entrenamiento (15 frases grabadas con 2 micrófonos) y 4 locuciones de test (2 frases grabadas con 2 micrófonos).

Con la división descrita, los conjuntos no son independientes del locutor. Aunque a priori parece interesante conseguir unos modelos independientes del locutor, debido a que este clasificador va a ser utilizado en el reconocedor de habla, y en él tanto el conjunto de entrenamiento como el de test contiene a todos los locutores, con ninguna de las dos divisiones propuestas se consigue este propósito. Para que todos los locutores estén en igualdad de condiciones, en el reconocimiento se utilizará el clasificador de género diseñado con la actual división de datos.

En primer lugar se hace una batería de pruebas para ver el efecto del número de gausianas utilizadas en los GMMs. En todos estos experimentos el umbral utilizado será '0'. Por las conclusiones obtenidas en las pruebas de la división de datos anterior, la reestimación que se realizará en cada incremento de las gausianas constará de 3 ejecuciones del algoritmo de Baum-Welch.

Una vez elegido el número de gausianas que produce mejores prestaciones, se ajustará el umbral, de forma que se igualen las tasas de acierto de los locutores de ambos géneros.

a) Ajuste al número de gausianas

En cuanto a la evaluación del número de gausianas de los modelos, teniendo en cuenta un valor 0 para el umbral, los resultados obtenidos se muestran en la Figura 25 (datos en la Tabla 14 del Anexo I).

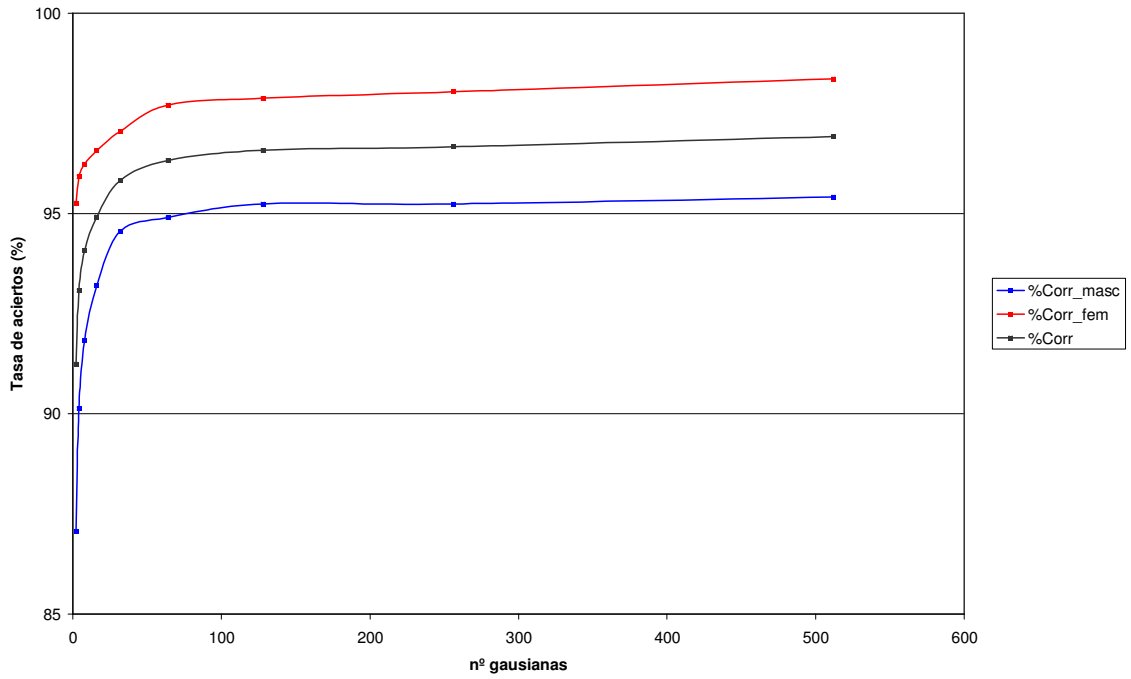

Figura 25.- Grupos dependientes del locutor – 3 reestimaciones

Como se puede observar, en estas pruebas no se tiene el mismo comportamiento que en el caso del conjunto anterior. La tasa de aciertos aumenta tanto para los locutores femeninos como para los masculinos a medida que aumenta el número de gausianas. Este hecho se debe estar produciendo porque se entrena y se testea con el mismo conjunto de locutores, y al aumentar el número de gausianas los modelos de género se están adaptando cada vez más a los locutores, por lo que con cada incremento se obtienen mejores resultados. A pesar de este efecto, existe un punto a partir del cual el aumento de la tasa de aciertos es cada vez más lento, creciendo muy lentamente a partir de 64 gausianas.

Por el mismo motivo comentado en casos anteriores, reducir la complejidad de la fase de test del clasificador de género, se considera adecuada la elección de 64 gausianas para esta división de los datos.

b) Ajuste del umbral

En la Figura 25 se observa que la tasa de aciertos femeninos es siempre mayor que la de los locutores masculinos. Para igualar los resultados obtenidos con ambos géneros se pasa a realizar un ajuste del umbral de decisión. En la Figura 26 (datos en la Tabla 15 del Anexo I) se muestran los resultados obtenidos.

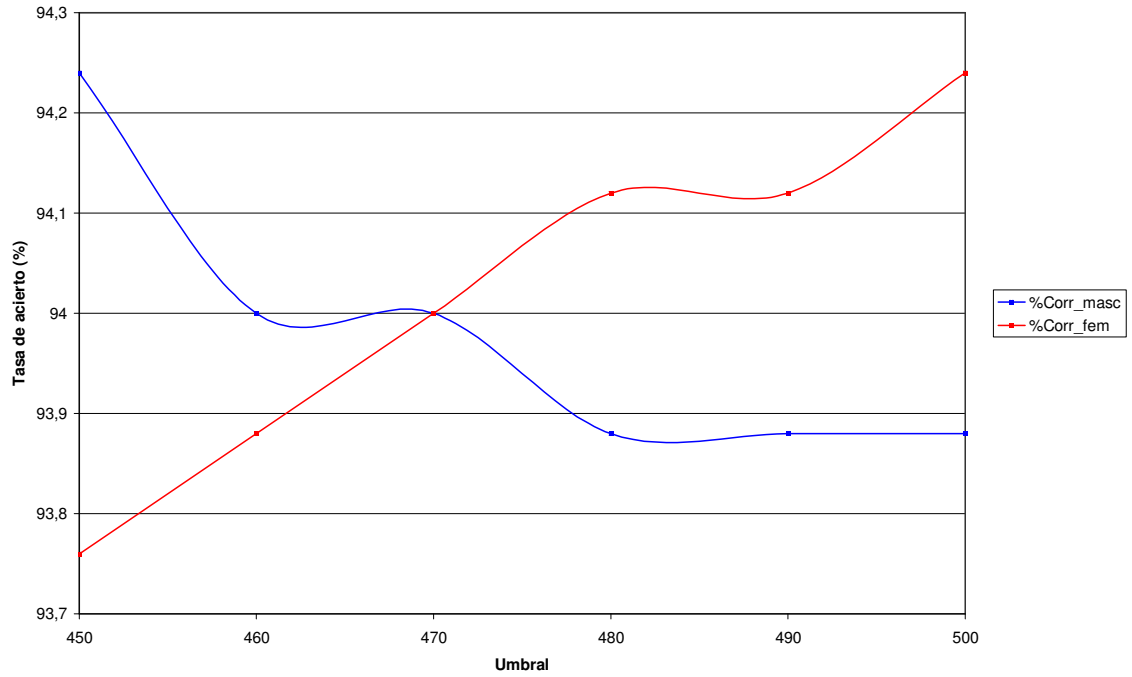

Figura 26.- Grupos dependientes del locutor – ajuste del umbral

El umbral óptimo en este caso es de 470, valor para el que la tasa de aciertos masculinos es igual a la de femeninos, teniendo ambas un valor de 94.00%.

4.3.4 Conclusiones

De los experimentos realizados sobre el clasificador de género, se sacan las siguientes conclusiones:

- El aumento del número de gausianas tiene influencia positiva en los resultados del clasificador. Existe un punto a partir del cual el incremento de las prestaciones del clasificador es muy pequeño, e incluso podría empezar a disminuir, ya que puede estar produciéndose una sobreadaptación a los datos de entrenamiento.
- El número de reestimaciones entre incrementos de gausianas también influye positivamente en la calidad de los modelos, llegándose a mejores resultados con menor número de gausianas y mayor número de reestimaciones.
- Debido a que el número de gausianas tiene una repercusión en el tiempo necesario para decidir el género de una locución de entrada, se debe elegir el modelado que produzca mejores resultados para un menor valor del número de gausianas. El tiempo necesario para el entrenamiento se considera un factor secundario, y por lo tanto el número de reestimaciones necesarias.
- En cuanto al ajuste del umbral, éste se ha realizado de forma experimental para los datos disponibles, decidiéndose por el valor que aporta una igualdad en la tasa de acierto de ambos géneros. Esto hará que para ciertas locuciones se pueda elegir un género equivocado. Lo que para el clasificador de género es un error, puede hacer que en reconocimiento aporte beneficio, ya que esa locución podría tener características de una locución femenina, o viceversa.
- Se han realizado dos baterías de pruebas para dos divisiones diferentes de los datos. Queda demostrado que en los modelos, además del género, se modela cierta información del locutor, ya que los resultados aumentan al utilizar el grupo dependiente del locutor.
- A pesar de lo anterior, el conjunto de datos elegido para entrenar el clasificador de género utilizado finalmente en el reconocedor es el dependiente del locutor. Se eligió esta división para que todos los locutores estuvieran en igualdad de condiciones, ya que en el reconocimiento, tanto el grupo de entrenamiento como el de test contienen la totalidad de los locutores.
- Los modelos elegidos finalmente son 3 GMMs diferentes, modelo 'mundo' (λ_{mundo}), modelo masculino (λ_{masc}) y modelo femenino (λ_{fem}), formados por una mezcla de 64 gausianas. Para el entrenamiento del modelo 'mundo' se utiliza una reestimación

con 3 realizaciones del algoritmo de Baum-Welch en cada incremento del número de gausianas. Los modelos de género se obtienen por adaptación MAP del modelo anterior, y para la clasificación se utilizará un umbral igual a 470.

4.4 Experimentos realizados

En este apartado se van a explicar las distintas pruebas realizadas sobre el reconocedor para evaluar la influencia del género de los locutores en el reconocimiento. Cada una de las modificaciones realizadas sobre el reconocedor de partida será explicada en el apartado correspondiente.

4.4.1 Adaptación MAP al género

En este primer grupo de experimentos se pretende obtener dos conjuntos de modelos (correspondientes a trifonemas, más los modelos de las pausas entre palabras) adaptados al género, uno al género masculino y el otro al género femenino, partiendo de unos modelos globales bien entrenados. La técnica de adaptación utilizada es la adaptación MAP.

Como esta adaptación no está contemplada en el reconocedor inicial, éste debe ser modificado para utilizarla.

Las nuevas etapas introducidas en la fase de entrenamiento del reconocedor, una vez obtenido el conjunto de modelos genéricos, son las siguientes:

- Agrupamiento de los datos de entrenamiento en función del género de los locutores. De esta forma se obtendrán dos conjuntos de datos, denominados datos de adaptación, uno con las locuciones masculinas y otro con las femeninas.
- Realización de la adaptación del conjunto de modelos genéricos a las locuciones masculinas. HTK ofrece una herramienta que realiza dicha adaptación, a la que se le indica la parte de los modelos que se pretende adaptar: media, varianza, y/o matriz de transición. Como ya se ha comentado, HTK no permite adaptar los pesos de las gausianas.
- Realización de la adaptación del conjunto de modelos genéricos a las locuciones femeninas.

En la Figura 27 se muestra esquemáticamente la modificación de la fase de entrenamiento del reconocedor.

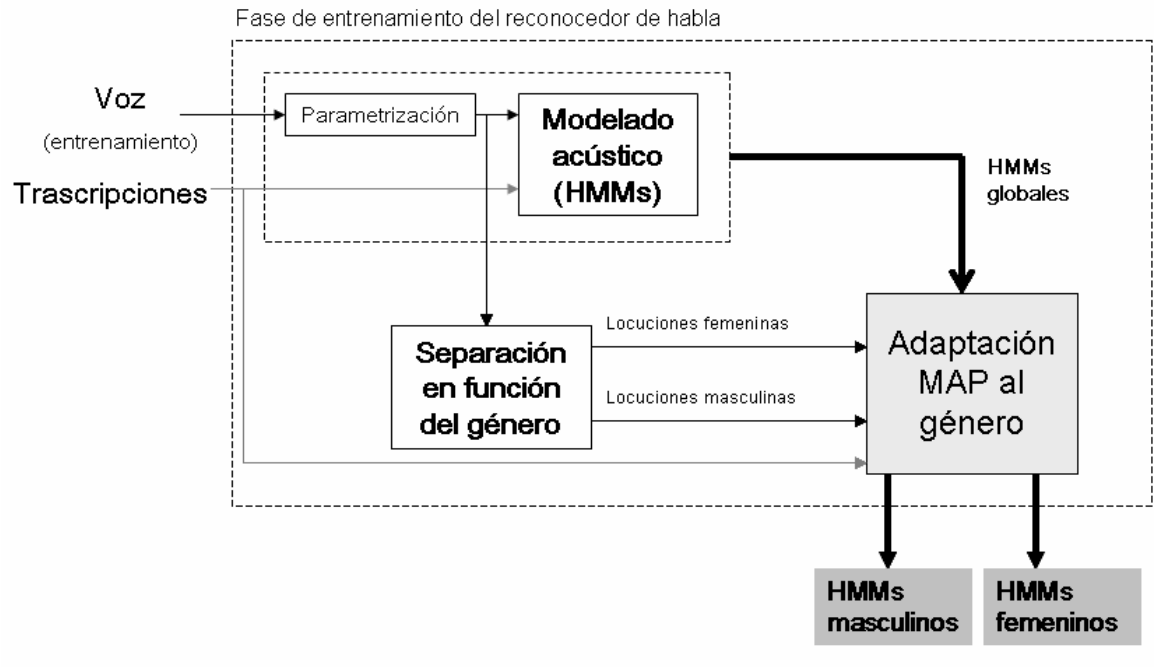

Figura 27.- Entrenamiento del reconocedor de habla con adaptación al género

En cuanto a la separación de los datos de entrenamiento para formar los dos conjuntos de datos de adaptación, se puede realizar de dos formas:

- Atendiendo al género real de los locutores, que es una información ofrecida por la propia base de datos.
- En función del género que el clasificador de género indica para cada locución. De esta manera, locuciones de un género cuyas características están más cercanas a las del género contrario serán utilizadas para la adaptación del conjunto de modelos de dicho género.

Una vez se han obtenido los dos conjuntos de HMMs adaptados al género, la fase de reconocimiento también debe ser modificada, de forma que en función del género de la locución bajo test se elija un conjunto u otro de modelos para realizar el reconocimiento. Esta modificación queda representada esquemáticamente en Figura 28.

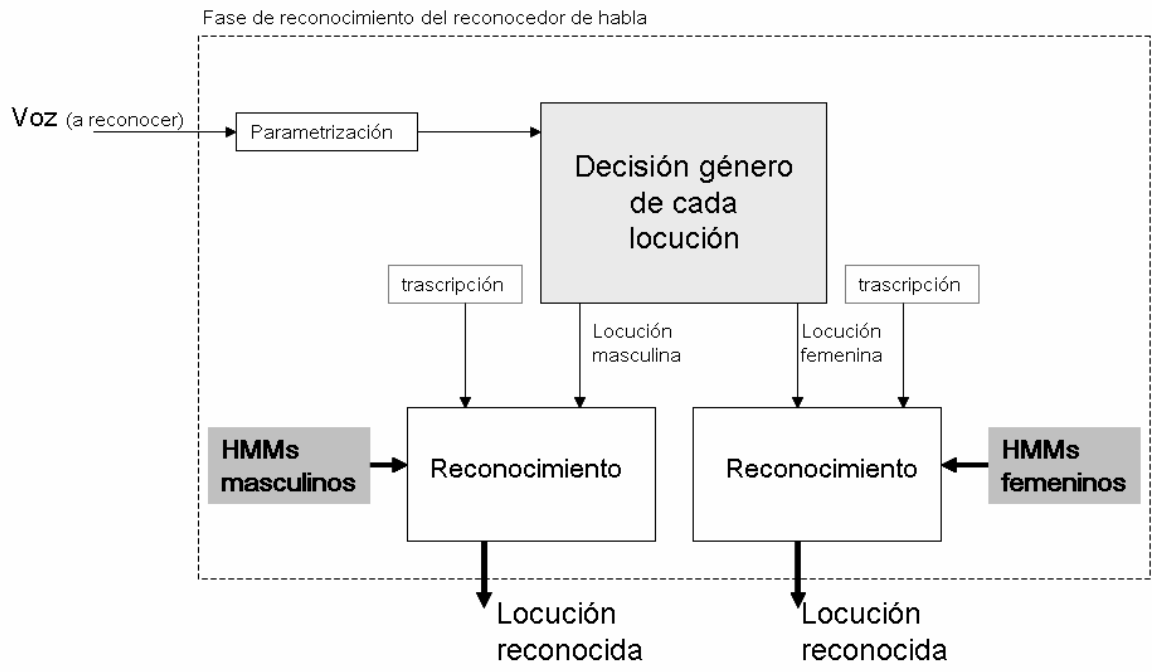

Figura 28.-Reconocimiento con modelos adaptados al género

De nuevo es necesario obtener el género de la locución, pudiéndose utilizar bien el género real, indicado por la base de datos, o bien el género decidido por el clasificador, lo que en principio parece una aplicación más real, ya que esta información puede no ser conocida.

Una vez realizado el reconocimiento de todas las locuciones de test, e igual que en el reconocedor inicial, se obtendrá una serie de medidas para poder evaluar los resultados.

Hasta aquí se han descrito las modificaciones llevadas a cabo en el reconocedor inicial para introducir la adaptación al género. A continuación se muestran los distintos experimentos realizados.

4.4.1.1 Utilización del género real en la fase de entrenamiento

En esta primera batería de pruebas realizadas sobre adaptación MAP se va a utilizar el género real de los datos de entrenamiento para obtener los dos conjuntos de datos de adaptación. Se realizan pruebas de adaptación de las medias y de adaptación de las medias y varianzas, además de un barrido de τ para observar la influencia del valor de este parámetro en el proceso de adaptación.

a) Adaptación MAP de las medias

Para los experimentos aquí realizados, los modelos finales se obtienen mediante adaptación MAP de las medias, es decir, las varianzas, pesos de las gausianas y matrices de transición son comunes en ambos grupos de modelos e iguales a los del conjunto de modelos genéricos.

Una vez obtenidos los HMMs adaptados, se procede a su evaluación, utilizando para ello el género real de las locuciones de test. Los resultados obtenidos se muestran a continuación (Resultado2):

Acc=73.94% [H=54787, D=5088, S=11037, I=2356, N=70912]

Resultado2.- Adaptación MAP de las medias con género real, test con género real

Con respecto al experimento base (Resultado1), sin utilizar información sobre el género de los locutores, se ha conseguido un aumento de la probabilidad de acierto de 0.82%, con un intervalo de confianza de ±0,32% para una confianza del 95%. Con estos valores y sabiendo que el intervalo de confianza para el experimento base es de ±0.33%, se puede decir que la mejora obtenida es estadísticamente significativa, ya que las bandas de confianza no se solapan. En cuanto a la mejora de la tasa de error relativa, sabiendo que para el experimento base se tiene una tasa de error de 26.88%, se está produciendo una mejora relativa del 3.05%.

La utilización del género real de las locuciones en la fase de test limita la aplicación del reconocedor diseñado a entornos donde esta información esté disponible. Ya que esta información puede no conocerse, se decidió utilizar el clasificador de género para obtener el género de las locuciones de test, y de esta forma el reconocedor puede utilizarse en cualquier entorno.

Utilizando el género indicado por el clasificador para las locuciones de test los resultados obtenidos son (Resultado3):

Acc=73.93% [H=54779, D=5095, S=11394, I=2410, N=70912]

Resultado3.- Adaptación MAP de las medias con género real, test con género del clasificador

Los resultados son muy parecidos al caso anterior, produciéndose un ligero empeoramiento (pasando a una mejora relativa del error de 3.01%) provocado por la tasa de error del clasificador de género, que hace que locuciones de un género utilicen los modelos del género opuesto para el reconocimiento.

b) Adaptación MAP de las medias y varianzas

A continuación se evalúa el comportamiento producido al introducir la varianza en la adaptación de los modelos, obteniéndose (Resultado4):

Acc=73.97% [H=54911, D=4981, S=11020, I=2459, N=70912]

Resultado4.- Adaptación MAP de las medias y varianzas con género real, test con género del clasificador

Se tiene cierta mejora sobre los resultados producidos al adaptar sólo las medias, tratándose de una mejora relativa de la tasa de error de 0.15% respecto al caso de adaptar sólo las medias. Comparando la tasa de acierto para el caso de adaptar sólo las medias, Resultado3, donde se tiene 73.93 ± 0.32%, y el caso de adaptar medias y varianzas, Resultado4, donde se produce una tasa de acierto de 73.97 ± 0.32%, se puede concluir que la mejora producida al introducir las varianzas no es estadísticamente significativa, ya que se solapan los rangos indicados por los intervalos de confianza. Además, introducir las varianzas en el proceso de adaptación provoca un aumento importante en la complejidad del reconocedor.

Por último, en esta batería de pruebas se pretende evaluar la influencia del parámetro τ sobre los resultados, buscando un valor que provoque la máxima tasa de aciertos. Hasta ahora se ha estado utilizando el valor por defecto que ofrece HTK, siendo este de 10.

El barrido realizado queda representado en la Figura 29 (datos en la Tabla 16 del Anexo I). Hay que tener en cuenta que se adaptan tanto las medias como las varianzas.

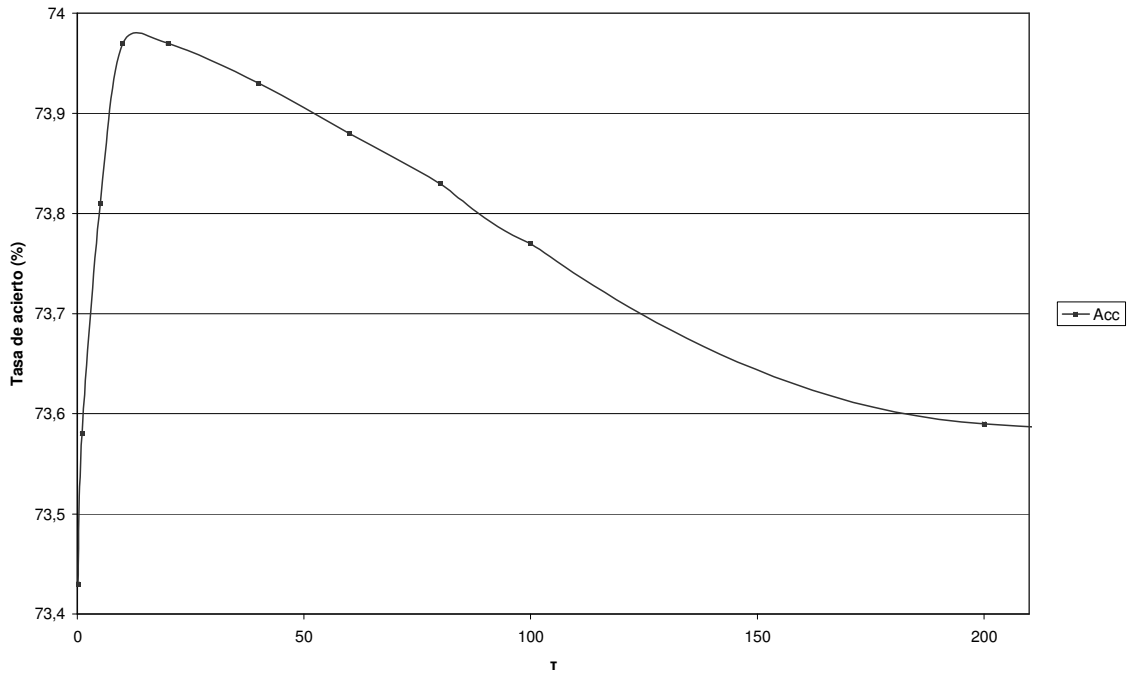

Figura 29.- Barrido de τ (adaptación con género real, reconocimiento con género del clasificador)

Como se puede ver los mejores resultados se obtienen para valore pequeños de τ. Esto puede interpretarse como que la cantidad de datos de adaptación existentes para cada género es muy elevada y, debido a esto, al aumentar la importancia que se le da a la información a priori, en detrimento de la influencia de la información que aportan los datos de adaptación, los resultados empeoran. Con esta información parece que se obtendrían buenos resultados si se entrenaran los modelos de género desde el principio, sin necesidad de realizar la adaptación.

4.4.1.2 Utilización del género decidido por el clasificador en la fase de entrenamiento

En las pruebas del apartado anterior se observa que al utilizar el género decidido por el clasificador en las locuciones de test se produce un pequeño empeoramiento de los resultados, pasando de una tasa de aciertos del 73.94% a una de 73.93%. Esto puede ser debido a que existen locuciones de test cuyas características de género se asemejan más a las del género opuesto, obteniéndose el género equivocado para dicha locución.

En esta batería de pruebas se pretende obtener beneficio de la observación anterior. Si una determinada locución se caracteriza mejor con el género contrario, esta locución debería intervenir en la adaptación del conjunto de modelos de dicho género. De forma contraria, esta

locución estaría introduciendo características erróneas en el modelado. Si después, en la fase de reconocimiento, se utiliza este mismo principio, utilizando el conjunto de modelos más afín a las características de la locución (aunque se trate en realidad de un conjunto equivocado), es previsible que los resultados de reconocimiento mejoren respecto a los obtenidos con el género real.

Utilizando el clasificador de género tanto en la fase de adaptación como en la de reconocimiento se obtienen los siguientes resultados (Resultado5):

Acc=74.10% [H=54871, D=5090, S=10951, I=2326, N=70912]

Resultado5.- Adaptación MAP de las medias con género del clasificador, test con género del clasificador

Comparando estos resultados con los del caso en el que se utiliza el género real para la adaptación y el del clasificador para el test (Resultado3), se ve que se produce un aumento de la tasa de aciertos de 0.17%. La mejora relativa de la tasa de error obtenida en este caso es de 3.64%, mientras que en el caso de no utilizar el clasificador en la fase de adaptación se tiene una mejora relativa de 3.01%. Aunque debido a los intervalos de confianza (de ± 0.32% en ambos casos), no se está consiguiendo una diferencia estadísticamente significativa, se obtiene una mejora de 0,63% en el error relativo, con lo que se toma como beneficioso el uso del género del clasificador tanto en la fase de entrenamiento como en la de test.

Para complementar el experimento anterior se realiza un barrido de τ, esperando ver un comportamiento semejante al obtenido en el barrido anterior. En este caso, sólo se adaptan las medias. Los resultados se muestran en la Figura 30 (datos en la Tabla 17 del Anexo I).

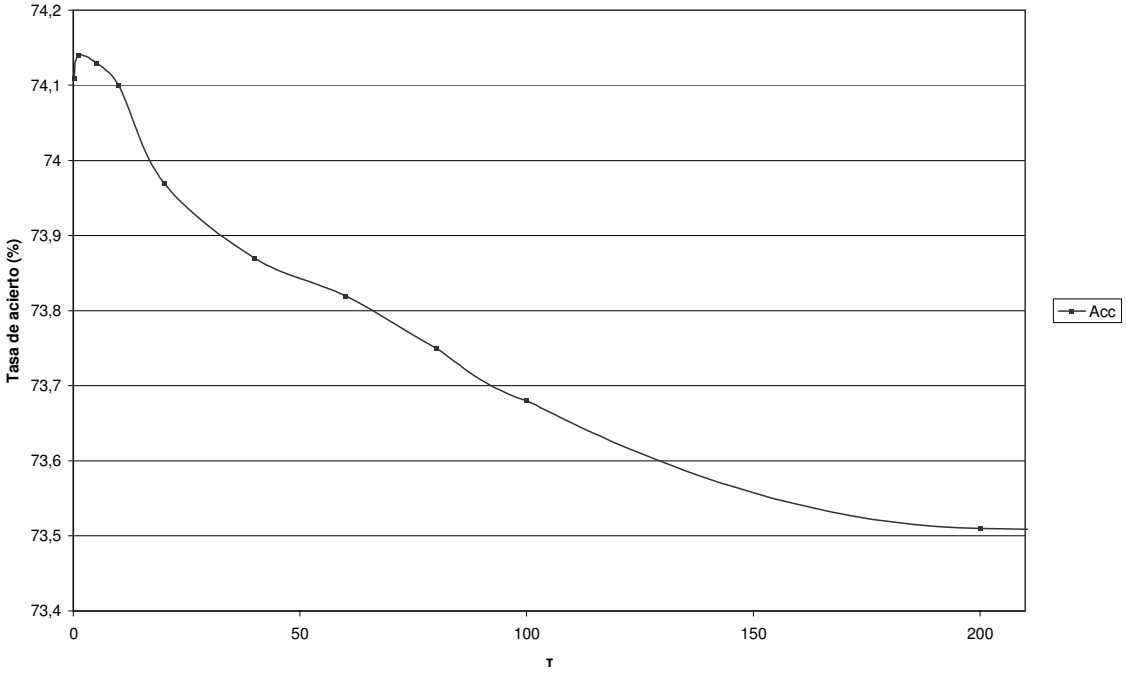

Figura 30.- Barrido de τ (adaptación y reconocimiento con género del clasificador)

Como se puede comprobar, el efecto es el mismo que en la Figura 29, los resultados mejoran para valores pequeños de τ , aunque para este caso los valores máximos conseguidos, a pesar de estar adaptando solamente las medias, son mayores que los obtenidos para el caso de utilizar el género real en la adaptación de los modelos.

4.4.2 Entrenamiento completo por género

Ya que existen indicios de que los datos de entrenamiento disponibles para cada género son bastante elevados, se pretende comprobar si esta cantidad es suficiente para realizar un entrenamiento completo de cada conjunto de modelos dependiente del género, sin necesidad de realizar la adaptación de un conjunto de HMMs entrenado con una cantidad de datos mayor. El esquema correspondiente a la fase de entrenamiento del reconocedor se correspondería con la Figura 31.

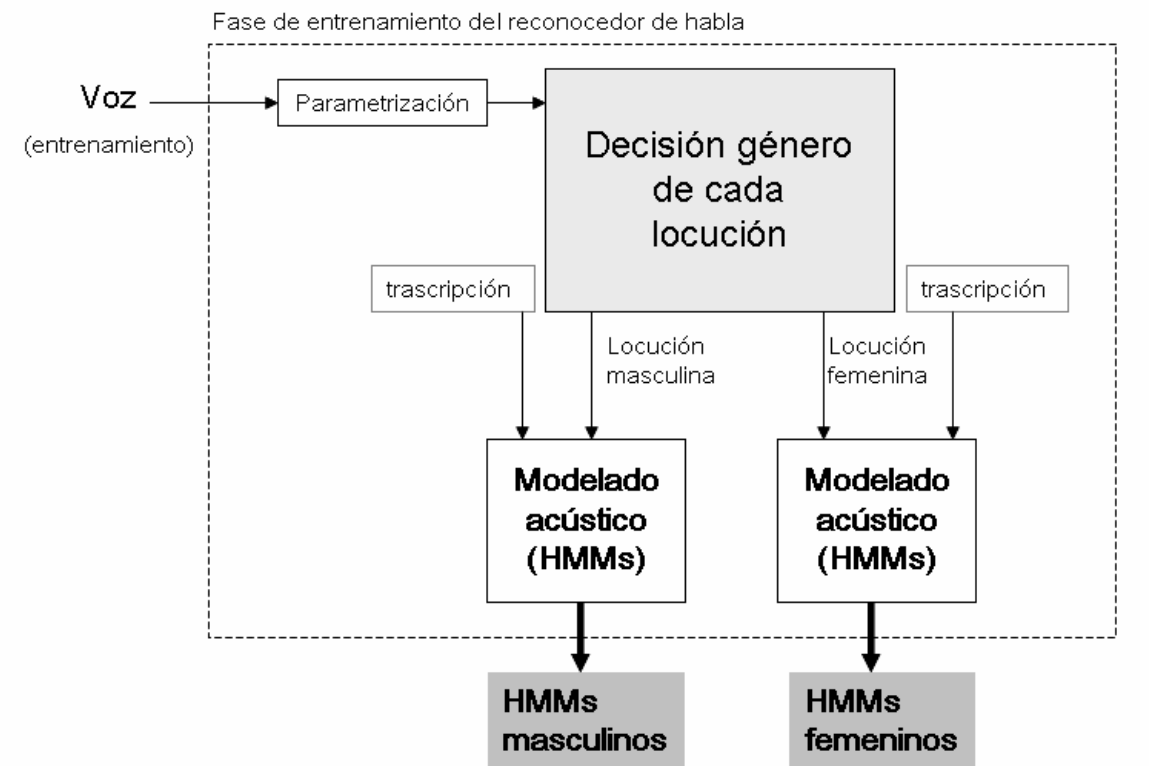

Figura 31.- Entrenamiento del reconocedor de habla con modelos separados por género

Utilizando la separación de los datos de entrenamiento según la información disponible en la base de datos, es decir, según el género real, y la separación de los datos de test según el clasificador, se obtiene (Resultado6):

Acc=74.61% [H=55389, D=4723, S=10800, I=2484, N=70912]

Resultado6.- Entrenamiento completo con género real. Reconocimiento con género del clasificador.

Con respecto al experimento utilizando adaptación MAP con esta misma división de datos (Resultado3), se está pasando de una tasa de aciertos de 73.93 ± 0.32 % a 74.61 ± 0.32 %, y teniendo en cuenta la mejora relativa de la tasa de error se está pasando de 3.01% a 5.54%. Teniendo en cuenta que se trata de unos resultados estadísticamente significativos, esto indica que, efectivamente, la cantidad de datos de entrenamiento es lo suficientemente alta como para que no sea necesario utilizar técnicas de adaptación.

Introduciendo en el entrenamiento la información aportada por el clasificador de género, que ya se ha comprobado que en adaptación produce una mejora de los resultados, se obtiene:

Acc=74.89% [H=55531, D=4798, S=10583, I=2426, N=70912]

Resultado7.- Entrenamiento completo y reconocimiento con género del clasificador.

Con respecto al experimento anterior (Resultado6), donde se utiliza el género real en la fase de entrenamiento y el género del clasificador en la fase de test, se está pasando de una tasa de aciertos de 74.61 ± 0.32 % a 74.89 ± 0.32 %, obteniendo una mejora relativa de la tasa de error de 6.58% respecto al experimento base. Igual que sucedió en el caso de utilizar adaptación, aunque debido a los intervalos de confianza (de ± 0.32% en ambos casos) no se está consiguiendo una diferencia estadísticamente significativa entre los experimentos mostrados como Resultado6 y Resultado7, se está obteniendo una mejora de 1.04 % en el error relativo, por lo que se concluye que el uso del género del clasificador tanto en la fase de entrenamiento como en la de test es beneficioso respecto al caso de utilizar el género real.

4.5 Conclusiones

La principal conclusión obtenida de los experimentos realizados es que diferenciar la generación de los modelos del reconocedor en función del género de los locutores aporta beneficios en el proceso de reconocimiento. Con esto queda comprobado que las características de la señal de voz tienen dependencia con el género.

En cuanto a la adaptación MAP, indicar que se está consiguiendo un valor de 74.14% para la tasa de aciertos, utilizando un valor de τ igual a 1, adaptando sólo las medias, y utilizando el clasificador de género tanto para la elección de los datos de adaptación como de test. Este valor representa una mejora relativa de la tasa de error de 3.79% respecto al experimento base (Resultado1).

Aunque la mejora obtenida con la adaptación es bastante significativa, se ha demostrado que los datos de adaptación son lo suficientemente extensos como para obtenerse mejores resultados entrenando los modelos dependientes del género desde el principio, sin necesidad de realizar ningún tipo de adaptación a un conjunto de modelos genéricos. Utilizando el género real para la división de los datos de entrenamiento se obtiene una tasa de acierto de 74.61%, lo que equivale a una mejora relativa de la tasa de error de 5.54% respecto al experimento base (Resultado1). Si además, se utiliza el género del clasificador tanto en entrenamiento como en test, este incremento relativo asciende a 6.58%.

Utilizar el clasificador de género tanto en la fase de entrenamiento como en la de test provoca una mejora de los resultados, ya que permite aprovechar el hecho de que existan locuciones que presentan características más afines al género contrario. Utilizar esta información en entrenamiento, permite modelar cada género con muestras de similares características, que en caso contrario introducirían ruido en los modelos. El clasificador de género en la fase de test del reconocedor permite utilizar los modelos que mejor se corresponden con las características acústicas de las muestras de voz de la locución, lo que producirá transcripciones más exactas.

Capítulo 5

Efectos del locutor sobre el reconocimiento de habla

5.1 Resumen

En este capítulo se pretende observar en qué medida se ven afectados los resultados de reconocimiento de habla si se utiliza información sobre el locutor tanto en la fase de entrenamiento como en la de test.

Como hipótesis de partida, se considera que la información sobre el locutor que genera cada locución, tanto de entrenamiento como de test, es conocida. Con dicha información, se plantea la obtención de un conjunto de HMMs particular de cada locutor, para modelar los distintos trifonemas y los modelos de pausas.

Igual que en el caso de la utilización del género en el reconocedor, la obtención de estos nuevos conjuntos de modelos se plantea de dos formas diferentes:

- Realizando un entrenamiento completo, pero sólo con los datos del locutor particular para el que se están obteniendo los modelos. En cuanto a los datos de entrenamiento disponibles, se está pasando de 9000 locuciones a solamente 30, lo que a priori parece una cantidad de datos insuficientes para un buen modelado de todas las características de los HMMs.
- Debido a la limitación en el número de datos de entrenamiento, se plantea el uso de adaptación sobre el conjunto de modelos ya obtenido, que estará suficientemente bien entrenado puesto que se ha usado el conjunto completo de datos para su generación. Se plantean dos técnicas de adaptación: adaptación MAP y adaptación MLLR (*Maximum Likelihood Linear Regression*). La primera de ellas [HAH01] aporta buenos resultados cuando la cantidad de datos de adaptación es suficientemente elevada, mientras que la adaptación MLLR responde mejor que la primera cuando se dispone de pocos datos de adaptación.

La técnica de adaptación MAP ha sido descrita en el capítulo 4 (4.2). Tras una introducción teórica sobre la técnica de adaptación MLLR, se explicará la batería de pruebas realizadas sobre el reconocedor para obtener información sobre la modificación producida en los resultados al tener en cuenta la información del locutor.

5.2 Adaptación al locutor

Como ya se ha comentado, dado que la base de datos consta de 300 locutores distintos, se pretende conseguir 300 grupos de modelos diferentes, caracterizando cada uno de ellos a un locutor determinado. Entrenar estos modelos desde el principio, utilizando únicamente los datos de entrenamiento disponibles de cada uno, presenta un gran inconveniente en cuanto a la gran cantidad de datos necesaria para el entrenamiento.

Igual que en el caso de adaptación al género, para solucionar el inconveniente que puede presentar la escasez de datos se plantea el uso de la adaptación [YEG+06], [HAH01]. Partiendo de un grupo de modelos inicial bien entrenado, se generan los nuevos modelos adaptando los primeros mediante los datos correspondientes a cada locutor. De esta forma son necesarios menos datos de adaptación, ya que se conoce en torno a qué valores se van a encontrar los parámetros de los modelos, y se trata únicamente de ajustarlos a cada locutor.

La cantidad de datos necesaria, y los resultados del reconocimiento realizado con estos modelos adaptados, dependerán del tipo de adaptación elegida. En este caso se van a utilizar dos técnicas diferentes, cuyas prestaciones dependen de la cantidad de datos disponibles.

5.2.1 Adaptación MAP

La adaptación MAP, en su versión supervisada, es una de las técnicas utilizada para la adaptación de los HMMs que se quiere realizar.

Esta técnica, descrita en el Apartado 4.2, realiza una transformación particular para cada modelo, siempre que existan datos de adaptación para ese modelo. Por lo tanto, cuanta más información se tenga disponible, mejores resultados ofrecerá.

5.2.2 Adaptación MLLR

La adaptación MLLR es una técnica utilizada en la adaptación de HMMs [Gal98], [HAH01], basada en el uso de transformaciones lineales. Se ha demostrado que estas transformaciones lineales presentan buenas características en la adaptación de las medias y varianzas de gausianas, por lo que pueden aplicarse fácilmente para la adaptación de HMMs, cuando las probabilidades de emisión de los distintos estados que los componen se modelan mediante mezclas de gausianas.

Por lo tanto, la adaptación MLLR trata de adaptar las medias y varianzas de dichas gausianas con pocos datos de adaptación, utilizando para ello funciones de transformación lineales, de forma que se maximice la probabilidad de los datos de adaptación dado el conjunto de modelos (criterio de *Maximum Likelihood*, ML), utilizando el algoritmo EM (*Expectation-Maximisation*) para resolver el problema de la maximización.

El efecto conseguido mediante la adaptación MLLR es un desplazamiento de las medias y una transformación de las varianzas, de forma que se reduzca la distancia existente entre los modelos originales y los datos de adaptación.

Aunque se puede adaptar tanto la media como la varianza, en los desarrollos implementados en este proyecto fin de carrera sólo se utiliza la adaptación de las medias, por lo que es obviada la forma de adaptación de las varianzas [YEG+06].

Al aplicar MLLR sobre el vector media de una gausiana que forma parte de un HMM, esta media se verá modificada de la manera indicada en (5.1):

$$\hat{\mu} = W\xi \qquad (5.1)$$

Siendo $\hat{\mu}$ la media estimada, que se obtiene aplicando la transformación fijada por la matriz W sobre un vector extendido de la media original, ξ, cuya expresión se muestra en *(5.2)*:

$$\xi = [w \quad \mu_1 \quad \ldots \quad \mu_n]^T \qquad (5.2)$$

Siendo w un factor de bias cuyo valor habitual es 1, μ_x las componentes del vector media original, cuya longitud es n. Para que con (5.1) se obtenga un vector con esta longitud n, la matriz de transformación, W, debe tener una dimensión nx(n+1). Esta característica permite ofrecer W descompuesta en los elementos indicados en (5.3):

$$W = [b \quad A] \qquad (5.3)$$

Conociéndose b como el vector bias y A como la matriz de transformación de dimensión nxn.

Esta transformación calculada puede ser compartida por diferentes gausianas, incluso por diferentes modelos. Para esto se generan una serie de árboles de regresión, que indican qué grupos de gausianas comparten las transformaciones obtenidas. Si la cantidad de datos de adaptación es escasa, es recomendable realizar una única transformación para todas las gausianas existentes en el conjunto de HMMs, de forma que todas se vean modificadas por la influencia de los datos de adaptación. A medida que se tenga mayor cantidad de datos, se puede ir agrupando componentes gausianas de modelos que compartan características acústicas, de forma que se vean modificados de la misma forma, y de diferente manera a las de los modelos que no comparten dichas características. Si se llegaran a tener agrupaciones de un único modelo, se tendrían transformaciones diferentes para cada uno, compartiendo el comportamiento de la adaptación MAP, donde sólo se adaptan los modelos que tienen presencia en los datos de adaptación.

HTK permite utilizar MLLR de forma supervisada, es decir, utilizando la trascripción correcta de los datos para realizar la adaptación, y de forma no supervisada, realizando un reconocimiento de los datos de adaptación con los HMMs ya existentes, para así obtener las transcripciones.

Además, se puede utilizar una versión incremental no supervisada de MLLR, para la cual se fija un número ‘n’ de locuciones a reconocer antes de realizar la adaptación. Superado ese número, se utilizan las transcripciones obtenidas para adaptar mediante MLLR los HMMs existentes. A partir de este momento se utilizan los modelos adaptados para obtener la trascripción de las siguientes ‘n’ locuciones de adaptación, volviendo a repetirse el proceso anterior.

En las pruebas llevadas a cabo en este proyecto, se utilizará un único árbol para realizar la adaptación MLLR, compuesto por todas las mezclas de todos los modelos, y se realizarán ejecuciones de dicha técnica en su versión supervisada e incremental no supervisada.

5.3 Experimentos realizados

En este apartado se van a explicar las distintas pruebas realizadas sobre el reconocedor para evaluar la influencia de la información sobre el locutor en el reconocimiento. Cada una de las modificaciones realizadas sobre el reconocedor de partida será explicada en el apartado correspondiente.

5.3.1 Adaptación MAP al locutor

En este primer grupo de experimentos se pretende obtener 300 conjuntos de modelos adaptados a cada locutor particular de los 300 disponibles en la base de datos, partiendo de unos modelos globales bien entrenados. La técnica utilizada es la adaptación MAP, en su versión supervisada, lo que indica que se dispone de la trascripción correcta de los datos de adaptación de cada locutor.

Al igual que tuvo que hacerse para introducir la adaptación MAP al género, el reconocedor inicial debe ser modificado para utilizar la adaptación MAP al locutor.

Las nuevas etapas introducidas en la fase de entrenamiento del reconocedor, una vez obtenido el conjunto de modelos genéricos, son las siguientes:

- Agrupamiento de los datos de entrenamiento en función del locutor para generar los datos de adaptación de cada uno de ellos.

- Para cada locutor, adaptación del conjunto de modelos genéricos a las locuciones del mismo.

Dichas modificaciones quedan representadas esquemáticamente en la Figura 32:

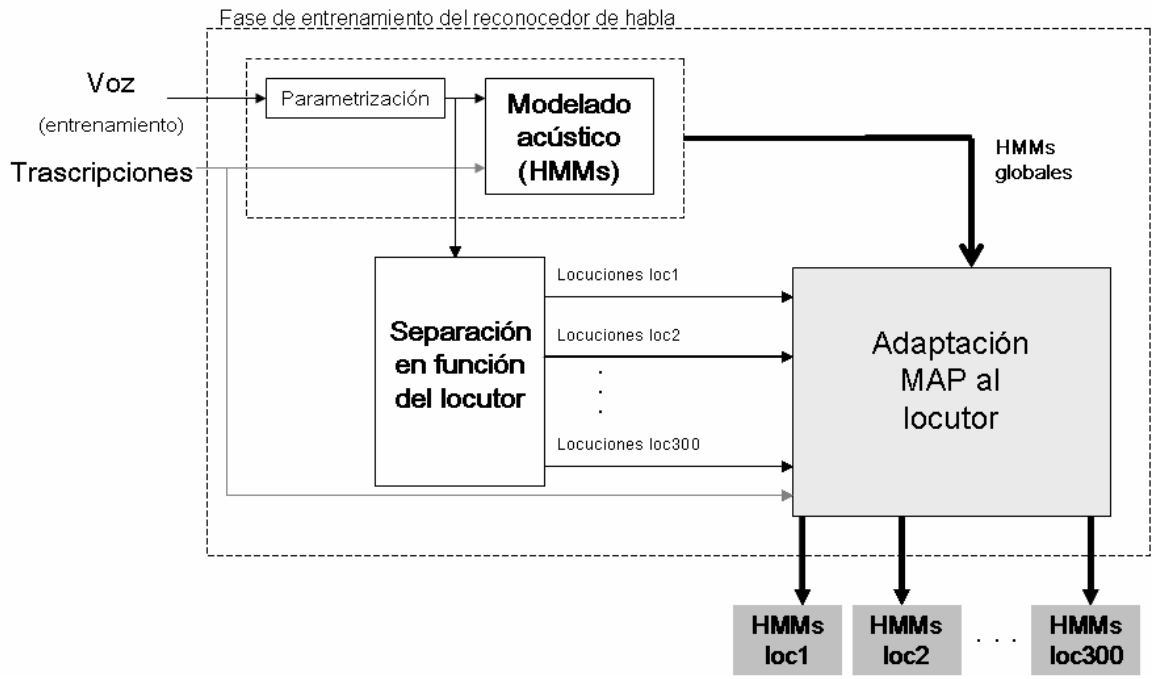

Figura 32.- Entrenamiento del reconocedor de habla con adaptación MAP al locutor

En cuanto a la separación de los datos de entrenamiento para formar los datos de adaptación, se puede realizar gracias a que la base de datos utilizada (MICROAES) indica, para todo archivo de voz disponible, el locutor que lo generó.

Una vez obtenidos los conjuntos de HMMs adaptados a cada locutor, la fase de reconocimiento también debe ser modificada (Figura 33), de forma que en función del locutor que generó la locución bajo test se elija el conjunto de modelos adaptado a dicho locutor para realizar el reconocimiento.

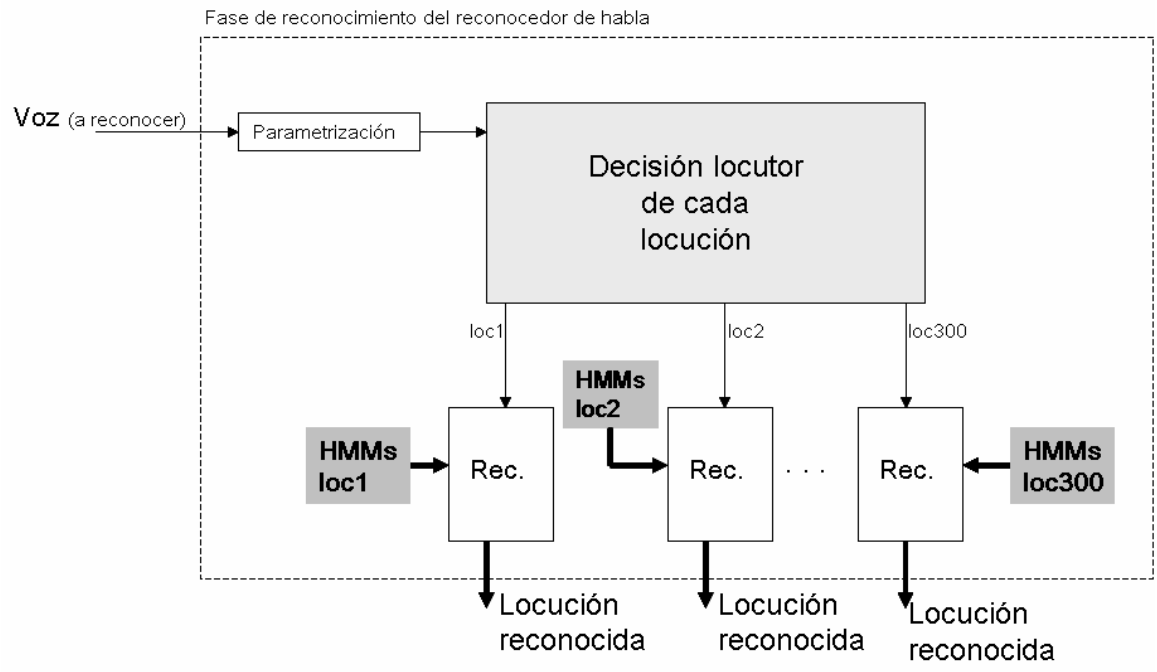

Figura 33.- Reconocimiento con modelos adaptados al locutor

Tras el reconocimiento de todas las locuciones de test, e igual que en todas las implementaciones del reconocedor vistas hasta ahora, se obtendrá una serie de medidas para poder evaluar los resultados.

Como inconveniente presente en este nuevo desarrollo hay que notar que se está pasando de 1 a 300 conjuntos de HMMs. En cada uno de ellos se almacena la información de todos los modelos utilizados para ese locutor, lo que representa un gran incremento en la capacidad de almacenamiento necesaria para este tipo de experimentos, factor a tener en cuenta en la implantación de este software en entornos donde el número de locutores sea elevado.

Sobre el reconocedor descrito, aplicando adaptación MAP al locutor, se realiza una batería de pruebas para ver el efecto que produce introducir la información sobre el locutor en el reconocimiento de habla. Se han realizado diferentes pruebas modificando el valor del parámetro τ, con el objetivo de obtener los mejores resultados posibles producidos para este tipo de adaptación, quedando representados en la Figura 5.1 (datos en la Tabla 18 del Anexo I).

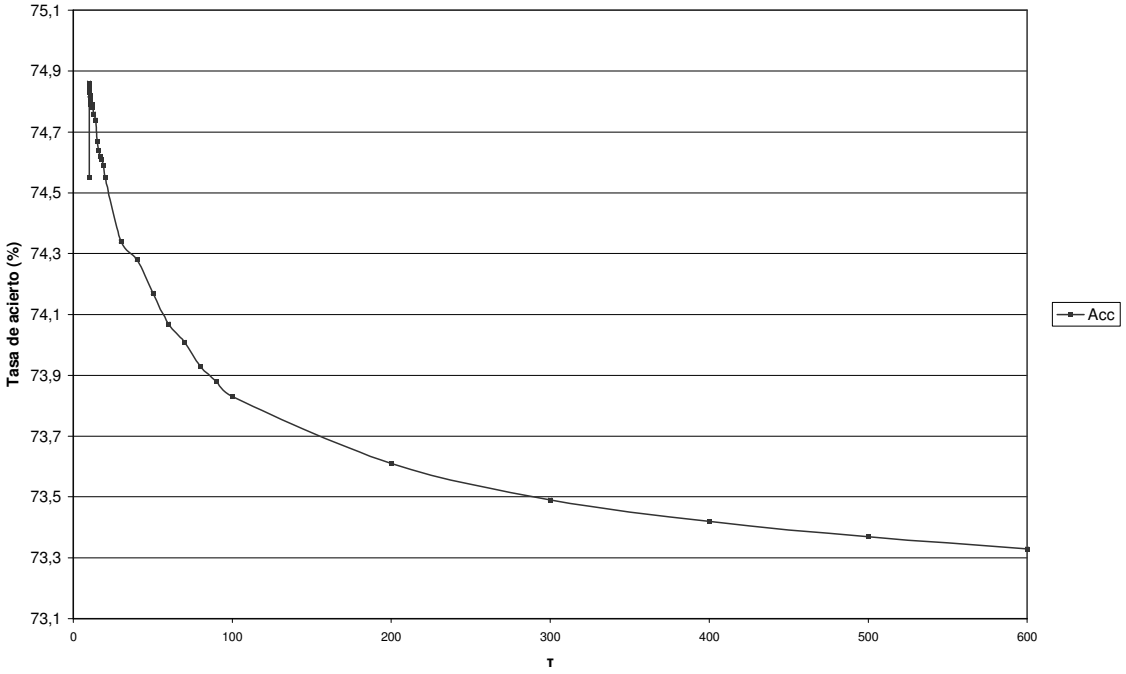

Figura 34.- Adaptación MAP al locutor (barrido de τ)

Como se puede observar, al introducir información sobre el locutor en los modelos de las distintas unidades acústicas utilizadas en el reconocimiento de habla, se produce una mejora significativa de los resultados de reconocimiento. Para un valor cercano a 10 del parámetro τ, más concretamente en un rango entre 10.1 y 10.5, se está consiguiendo una tasa de acierto de 74.86 ± 0.32%, lo que equivale a una mejora relativa de la tasa de error de 6.47% respecto al experimento base. Comparando los resultados con dicho experimento base (Resultado1), se está obteniendo un aumento de la tasa de acierto de 1.74%, suponiendo esto unos resultados estadísticamente significativos.

Con esto se puede concluir que la señal de voz transmite información representativa del locutor que la genera, y utilizar dicha información como complemento en el reconocedor de habla diseñado aporta mejoras significativas sobre los resultados.

5.3.2 Adaptación MLLR al locutor

Igual que para el caso de la adaptación MAP al locutor, se pretende realizar una adaptación a los datos del locutor utilizando adaptación MLLR. En principio esta segunda técnica aporta mejores resultados cuando existen pocos datos de adaptación, situación que parece ser la actual, ya que se tienen 30 locuciones de adaptación para cada locutor.

Además de lo anterior, mencionar que esta técnica aporta una ventaja respecto a la adaptación MAP. Como se ha comentado en 5.3.1, con la adaptación MAP se va a almacenar un nuevo conjunto de modelos para cada uno de los locutores, lo que representa un apreciable aumento en la capacidad de almacenamiento requerida. Para la adaptación MLLR, se tendrá un único conjunto de modelos, los correspondientes al entrenamiento genérico realizado, y para cada locutor adaptado sólo se almacenará la información necesaria para transformar esos modelos genéricos en los adaptados al locutor.

Se van a realizar diferentes pruebas sobre adaptación MLLR. Todas ellas comparten algunas características comunes, que se detallan a continuación:

- Sólo se realiza adaptación de las medias.
- Todas las medias sufren la misma transformación. Aunque la adaptación MLLR puede trabajar con agrupamientos de parámetros, de forma que se realice una adaptación diferente para cada grupo, en los experimentos aquí realizados se utiliza un único grupo, compuesto por todas las gausianas que componen los estados de todos los modelos del sistema.
- La transformación almacenada para cada locutor estará compuesta por el vector bias (b), de longitud 39, y la matriz de transformación (A), de dimensión 39x39.

La diferencia principal entre las pruebas realizadas estriba en que en la primera se utiliza una adaptación MLLR incremental, lo que quiere decir que se irá adaptando según se vayan obteniendo los datos de adaptación (versión no supervisada), mientras que la segunda se trata de una adaptación MLLR supervisada.

5.3.2.1 Adaptación MLLR incremental

La adaptación realizada es una adaptación no supervisada, partiendo de que no se dispone de la trascripción real de los datos de adaptación a cada locutor. Aunque la base de datos MICROAES adjunta la trascripción de todos los archivos de voz, incluidos los de test, se ha querido obviar aquí esta información para evaluar las prestaciones de este tipo de adaptación.

Además, se va a utilizar una adaptación incremental, lo que significa que según se vayan obteniendo los datos de adaptación se van a ir utilizando para adaptar los modelos, sin esperar a tener la totalidad de dicha información.

Para introducir esta adaptación en el reconocedor básico, es necesario modificar la fase de reconocimiento. La fase de entrenamiento concluye con la generación del conjunto de modelos genéricos. Sin embargo, el reconocimiento tendrá dos fases diferentes:

- Una primera fase (mostrada en la Figura 35) en la que, para cada locutor, se proceda al reconocimiento de las locuciones de adaptación de dicho locutor (locuciones obtenidas gracias a que la base de datos aporta información sobre el locutor), y con las transcripciones obtenidas se realiza la adaptación MLLR.
- Una segunda fase (mostrada en la Figura 36) para evaluar los resultados obtenidos de dicha adaptación, donde se reconocen las locuciones de test de cada locutor, utilizando para ello el conjunto de modelos genéricos con la transformación particular del locutor bajo test.

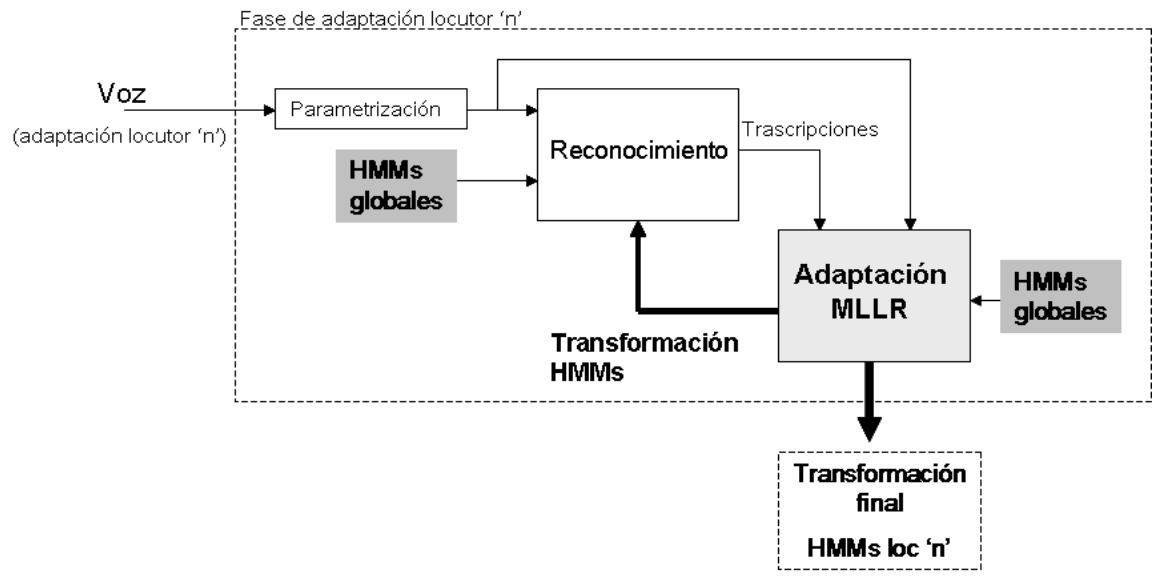

Figura 35.- Fase de adaptación para MLLR incremental no supervisada

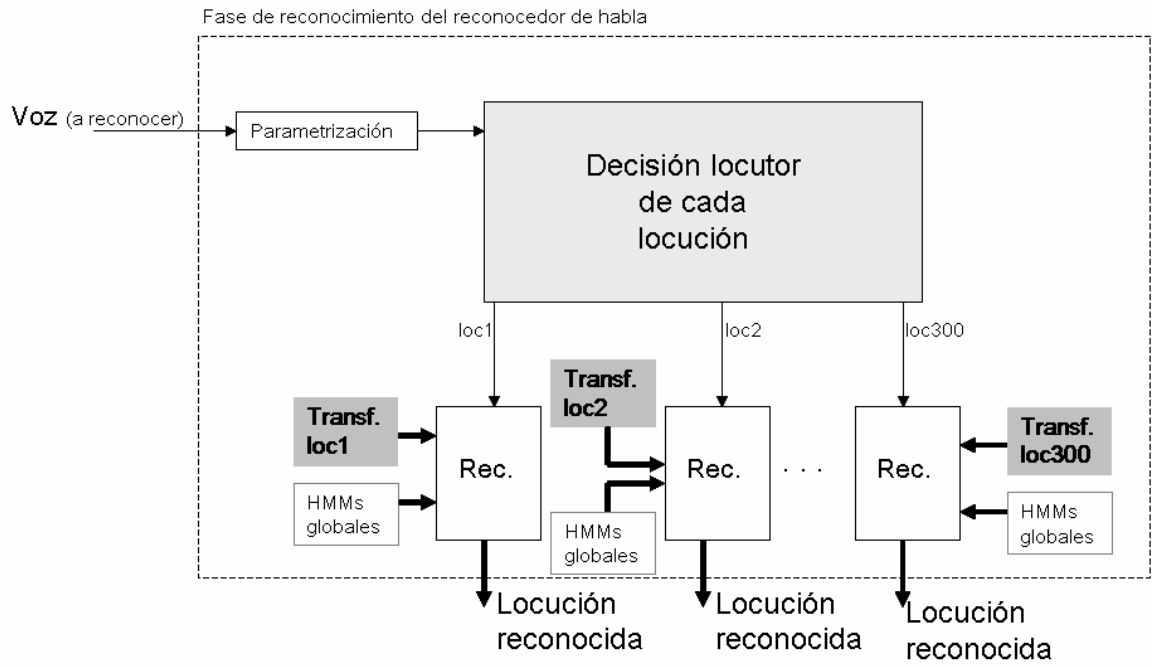

Figura 36.- Fase de reconocimiento para adaptación MLLR

La primera de las fases, al tratarse de una adaptación incremental, obtendrá la trascripción de ‘n’ locuciones de adaptación utilizando el conjunto de modelos genéricos, tras lo cual realizará la adaptación MLLR de dicho conjunto utilizando esas ‘n’ locuciones. Obtenida dicha transformación, utilizará el conjunto de modelos transformado para realizar el reconocimiento de las siguientes ‘n’ transcripciones, y así sucesivamente hasta el total de datos de adaptación.

Este tipo de adaptación permite que los modelos se adapten a la situación actual más o menos rápidamente, proporcionando mejores resultados que unos modelos estáticos.

Para poder realizar el reconocimiento de las locuciones de adaptación, ha sido necesario modificar el modelo de lenguaje del reconocedor, que sólo contenía las palabras de test. Se ha generado un nuevo modelo de lenguaje, utilizado sólo para la adaptación, con el conjunto de palabras que forman las locuciones de entrenamiento, un total de 7409 palabras.

Para evaluar los resultados de forma que puedan compararse las prestaciones con el resto de experimentos realizados, en la fase de reconocimiento se utiliza el modelo de lenguaje diseñado para los datos de test. Utilizando el nuevo modelo de lenguaje, se estaría eliminando la condición impuesta por el primero de que las palabras reconocidas siempre pertenezcan al conjunto limitado de palabras de test.

En la prueba realizada, se fija en 5 el número de locuciones que hay que reconocer para poder adaptar los modelos. Como cada locutor tiene 30 locuciones, se va a realizar un total de 6 adaptaciones incrementales para cada uno de ellos. Los resultados obtenidos, realizando el reconocimiento con la transformación final, se muestra a continuación (Resultado8):

Acc=76.08% [H=56147, D=4868, S=9897, I=2199, N=70912]

Resultado8.- Adaptación MLLR incremental al locutor

Los resultados muestran un aumento de la tasa de aciertos respecto a la adaptación MAP al locutor, pasando de 74,86 ± 0.32 % a 76,08 ± 0.31 %. Haciendo referencia a la mejora relativa de la tasa de error, se pasa de 6,47% a 11,01%.

De todo lo anterior se puede deducir que la adaptación al locutor aporta muy buenos resultados en cuanto a la tasa de aciertos del reconocedor de voz, y que la mejora introducida por la adaptación MLLR es significativa respecto a la adaptación MAP. Estos resultados eran previsibles, puesto que:

- Ya se había comprobado, mediante la adaptación MAP, que se producían mejoras de reconocimiento al introducir la información del locutor en los modelos.
- Los datos disponibles para cada locutor son muy limitados, y la adaptación MLLR ofrece mejores resultados que la adaptación MAP en este caso.

A pesar de la mejora en los resultados y a que los requisitos de espacio son mucho menores que con la adaptación MAP, es necesario indicar la existencia de un inconveniente que puede impedir el uso de la adaptación MLLR incremental en este entorno, y es el tiempo necesario para obtener la trascripción de las locuciones de adaptación. Si este tiempo es muy elevado, la capacidad de adaptación a la situación actual pierde importancia, ya que se tardará demasiado en adaptar los modelos a dicha situación. En el reconocedor desarrollado, el problema del aumento de dicho tiempo viene marcado por el modelo de lenguaje, ya que al aumentar el número de palabras, el número de combinaciones a evaluar es demasiado elevado.

5.3.2.2 Adaptación MLLR supervisada

Debido al aumento del tiempo necesario para realizar la adaptación MLLR incremental, y puesto que los datos con los que se está trabajando tienen disponible la trascripción de las locuciones de adaptación, se va a realizar la adaptación MLLR de forma supervisada, utilizando dicha información.

Para introducir esta adaptación en el reconocedor básico, se realiza de forma similar a la adaptación MAP (Figura 32). La fase de entrenamiento se verá modificada como sigue:

- Agrupamiento de los datos de entrenamiento en función del locutor para generar los datos de adaptación de cada uno de ellos.
- Para cada locutor, adaptación MLLR del conjunto de modelos genéricos a las locuciones del mismo.

El resultado de esta adaptación será la transformación a aplicar al conjunto de modelos genéricos para formar los modelos del locutor. Una vez obtenida la transformación correspondiente a cada uno de los locutores, en la fase de reconocimiento (Figura 36) se utilizará la transformación obtenida para el locutor bajo test para obtener los HMMs con los que realizar el reconocimiento.

Tras el reconocimiento de todas las locuciones de test, e igual que en todas las implementaciones del reconocedor vistas hasta ahora, se obtendrá una serie de medidas para poder evaluar los resultados.

Las prestaciones conseguidas con la adaptación MLLR implementada se muestran a continuación (Resultado9):

Acc=76.15% [H=56183, D=4825, S=9904, I=2185, N=70912]

Resultado9.- Adaptación MLLR supervisada al locutor

La tasa de aciertos obtenida supera ligeramente a la utilización de MLLR incremental, debido posiblemente al error introducido en la parte de reconocimiento de las transcripciones de adaptación.

Por lo tanto, se siguen consiguiendo resultados mejores que en el caso de adaptación MAP al locutor, pasando de 74,86 ± 0.32 % a 76,15% ± 0,31%. Haciendo referencia a la mejora relativa de la tasa de error, se pasa de una mejora de 6,47% a 11,27%.

Igual que para la prueba anterior, se puede concluir que la información del locutor introduce mejoras en los resultados de reconocimiento de habla, y que esta mejora aumento al utilizar adaptación MLLR en lugar de adaptación MAP. Además, al utilizar la trascripción real de las locuciones para la adaptación, las prestaciones mejoran.

5.3.3 Entrenamiento completo por locutor

Debido al hecho de que la adaptación MLLR aporta mejores resultados que la adaptación MAP, y puesto que la primera ofrece mejoras respecto a la segunda cuando la cantidad de datos de adaptación es escasa, parece patente que los datos disponibles de cada locutor no van a ser suficientes como para entrenar el conjunto de modelos desde el principio, con esos pocos datos de entrenamiento.

A pesar de esto, se ha realizado un entrenamiento, a modo de prueba, con un locutor al azar. Tras el reconocimiento utilizando el conjunto de HMMs obtenido, se obtiene una tasa de aciertos de 61.54% con un intervalo de confianza muy elevado, de 6,233%, debido a la poca cantidad de datos de test. A pesar de que el intervalo de confianza es grande, en todo caso la tasa de acierto de este locutor es inferior al valor del experimento de referencia (Resultado1), quedando demostrado así que los datos disponibles para cada locutor no son suficientes para realizar un buen entrenamiento.

5.3.4 Combinación de adaptación al género y al locutor

Puesto que ha quedado demostrado que la señal de voz contiene cierta información representativa del género y del locutor que la genera, y que la utilización, por separado, de ambas características provoca mejoras en los resultados de reconocimiento de habla, se plantea a continuación realizar una nueva modificación sobre el reconocedor base para utilizar una combinación de adaptación al género y adaptación al locutor.

Partiendo de unos modelos bien entrenados, generados con el conjunto completo de datos de entrenamiento, se obtendrán dos nuevos grupos de HMMs, cada uno de ellos resultado de la adaptación MAP al género del conjunto de modelos iniciales (Figura 27).

A continuación, para realizar la adaptación al locutor, en lugar de adaptar el conjunto de modelos genéricos, se adapta el conjunto de modelos correspondientes al género del locutor en cuestión (Figura 37). En este caso no se hace uso del clasificador de género, puesto que este clasificador indica el género que mejor se corresponde con cada locución, no el género del locutor al que corresponden una serie de locuciones. Debido a esto, se utilizará el género real para la elección del conjunto de modelos adaptado al género.

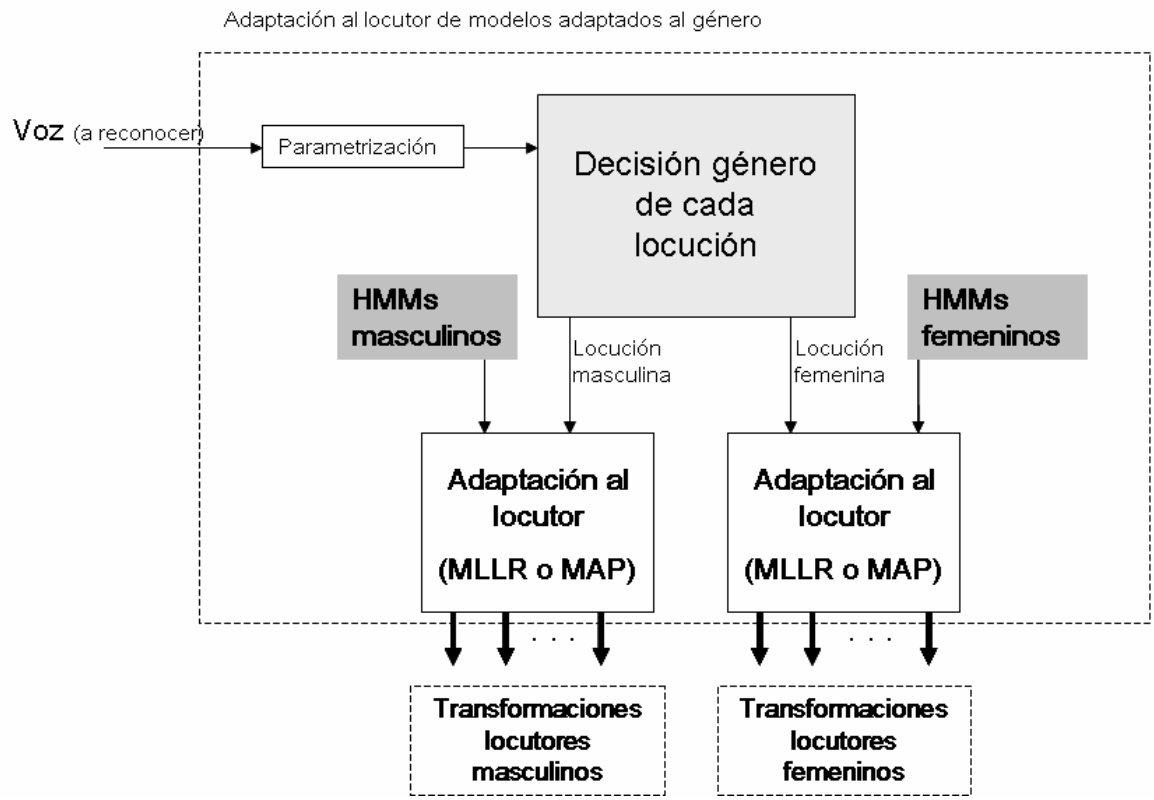

Figura 37.-Adaptación al locutor de modelos adaptados al género

La fase de reconocimiento debe ser modificada para utilizar los HMMs adecuados en función del locutor y el género del mismo.

Siguiendo estas directrices se realizan dos pruebas diferentes. En la primera de ellas se utiliza adaptación MAP tanto para la adaptación al género como para la adaptación al locutor, y en la segunda se plantea combinar la adaptación MAP para el género y la adaptación MLLR para el locutor.

5.3.4.1 Adaptación MAP al género y adaptación MAP al locutor

En esta primera combinación de adaptación al género y al locutor, se elige MAP como técnica a utilizar en ambos casos.

Como se ha comentado, el reconocedor debe ser modificado, tanto en el entrenamiento como en el reconocimiento. Las modificaciones realizadas en la fase de entrenamiento, tras obtener el conjunto de modelos inicial, se resumen a continuación:

- División de los datos de entrenamiento, en función del género indicado por el clasificador, para obtener los datos de adaptación al género. Se utiliza el clasificador partiendo de que con él se obtienen los mejores resultados en las pruebas realizadas en el capítulo 4.
- División de los datos de entrenamiento, en función del locutor al que pertenecen, para obtener los datos de adaptación al locutor.
- Adaptación MAP al género de los modelos originales, obteniendo dos nuevos conjuntos de modelos.
- Adaptación MAP al locutor del conjunto de HMMs correspondientes al género de dicho locutor.

En cuanto a la capacidad de almacenamiento requerida para estos experimentos, para cada locutor se obtiene un conjunto de modelos, adaptados tanto al género como al locutor, por lo que el tamaño ocupado cuando el número de locutores es alto puede ser muy elevado.

La fase de reconocimiento debe ser modificada para utilizar los HMMs adecuados en función del locutor y el género del mismo. En este caso, puesto que para cada locutor se tiene el conjunto de modelos adaptado tanto al género como al locutor, sólo será necesario conocer con qué locutor se está trabajando y utilizar el conjunto de modelos que le representa.

Para la adaptación al género se ha utilizado la mejor configuración obtenida hasta el momento, es decir, utilización del género decidido por el clasificador y un valor de τ igual a 1. En cuanto a la adaptación al locutor se ha realizado un barrido de τ para buscar los mejores

resultados. Este barrido se realiza en torno al valor 10, puesto que en la evaluación de la adaptación MAP al locutor llevada a cabo en 5.3.1 se vió cómo las mayores tasas de acierto se producía para τ cercanos a dicho valor.

Los resultados quedan representados en la Figura 38 (datos en la Tabla 19 del Anexo I).

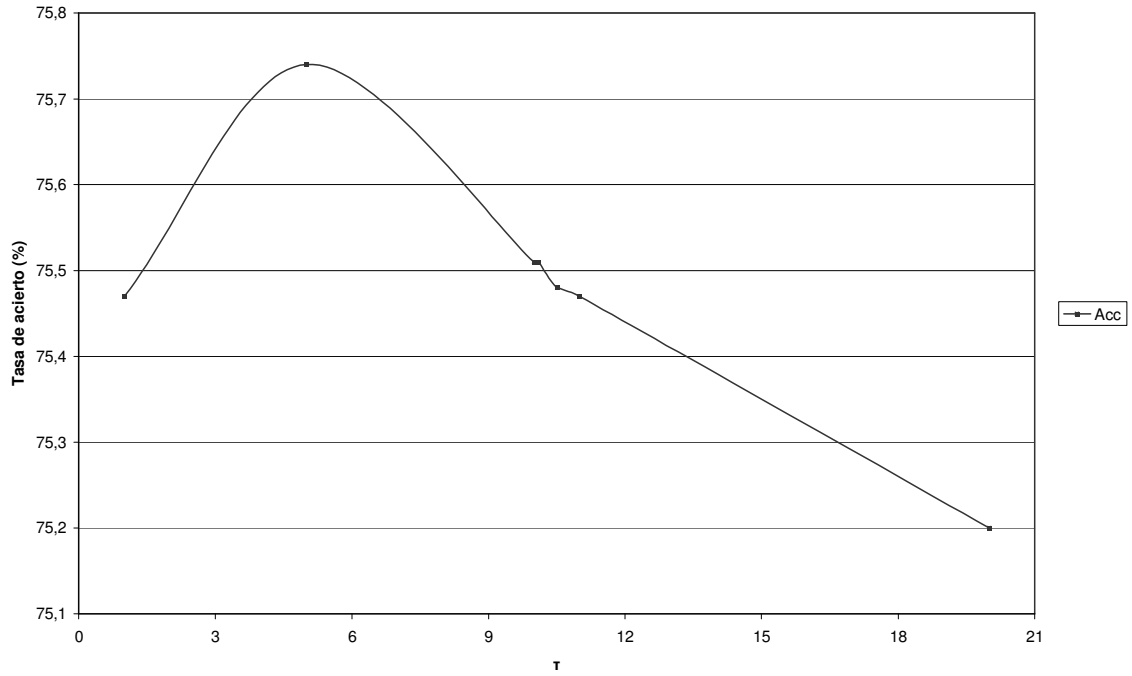

Figura 38.- Adaptación MAP al género y al locutor

Los resultados no sufren grandes cambios al modificar τ, consiguiéndose para un valor de 5 las mejores prestaciones, siendo la tasa de acierto de 75,74 ± 0,32 %. Comparando con el mejor caso de la adaptación al género, 74.14 ± 0,32 %, y con el mejor caso de la adaptación al locutor con MAP, 74.86 ± 0,32%, se comprueba que la combinación de ambas técnicas provoca un mayor valor en la tasa de aciertos que la aplicación por separado de ambas (con resultados estadísticamente significativos en los dos casos), consiguiéndose una mejora relativa de la tasa de error de 9.75% respecto al experimento base (Resultado1).

Esta combinación presenta los mejores valores conseguidos utilizando adaptación MAP, pero son peores a los obtenidos para el caso de adaptación al locutor con MLLR, por lo que se plantea combinar la adaptación MAP al género con la MLLR al locutor.

5.3.4.2 Adaptación MAP al género y adaptación MLLR al locutor

En esta segunda combinación de adaptación al género y al locutor, se elige MAP como técnica a utilizar en el primer caso, y MLLR, de forma supervisada, para la adaptación al locutor.

Igual que en el caso de adaptación MAP tanto al género como al locutor, el reconocedor debe ser modificado, tanto en el entrenamiento como en el reconocimiento. Las modificaciones realizadas en la fase de entrenamiento (Figura 27, Figura 37), tras obtener el conjunto de modelos inicial, se resumen a continuación:

- División de los datos de entrenamiento, en función del género indicado por el clasificador, para obtener los datos de adaptación al género.
- División de los datos de entrenamiento, en función del locutor al que pertenecen, para obtener los datos de adaptación al locutor.
- Adaptación MAP al género de los modelos originales, obteniendo dos nuevos conjuntos de modelos.
- Adaptación MLLR al locutor del conjunto de HMMs correspondientes al género de dicho locutor.

Como resultado al proceso anterior, se tendrán dos conjuntos de HMMs, cada uno de ellos adaptado a un género diferente, y la transformación a realizar sobre ellos (en función del género del locutor) para obtener el modelo adaptado al locutor y al género. Debido a esto, los requisitos de almacenamiento son menores que para el caso de utilizar MAP en ambas adaptaciones.

La fase de reconocimiento debe ser modificada para utilizar los HMMs adecuados en función del locutor y el género del mismo. En este caso, dado un locutor, y siendo conocido su género, se utilizará el conjunto de modelos adaptados a ese género, aplicándoles la transformación marcada por la adaptación MLLR al locutor de dicho conjunto de HMMs (Figura 39).

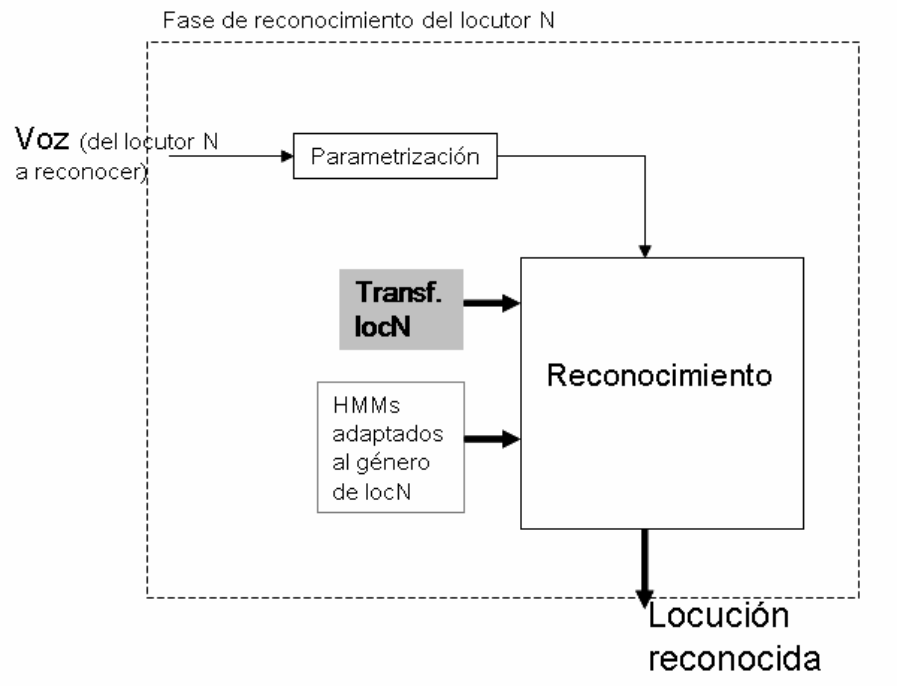

Figura 39.- Reconocimiento para locución del locutor 'N' con modelos adaptados al género con MAP y al locutor con MLLR

Tras el reconocimiento de todas las locuciones de test, e igual que en todas las implementaciones del reconocedor vistas hasta ahora, se obtendrá una serie de medidas para poder evaluar los resultados, que para este caso particular son (Resultado10):

Acc=76.73% [H=56608, D=4665, S=9639, I=2197, N=70912]

Resultado10.- Adaptación MAP al género y MLLR supervisada al locutor

Comparando con el mejor caso de la adaptación al género, 74.14 ± 0.32 %, y con el mejor caso de la adaptación al locutor con MLLR, 76.15 ± 0.31 %, se comprueba que la combinación de ambas técnicas provoca un mayor valor en la tasa de aciertos que la aplicación por separado de las mismas, consiguiéndose una mejora relativa de la tasa de error de 13.43% respecto al experimento base (Resultado1). Hasta ahora la mayor mejora relativa del error, 11.27%, se había conseguido para la adaptación MLLR al locutor, e introduciendo información sobre el género estamos aumentando este valor en un 2.16% más.

En cuanto a la comparación con el uso de MAP tanto para la adaptación al género como para la adaptación al locutor, se está pasando de una mejora relativa de 9.75% a 13.43%. Esto junto a la disminución de espacio ocupado al utilizar adaptación MAP al género y MLLR al locutor, lleva a decidirse por esta última implementación.

Se han obtenido unos resultados conforme a lo esperado, ya que teniendo en cuenta que se ha comprobado que la señal de voz transmite información relativa al género y al locutor, y que los datos disponibles separados por género son bastante abundantes mientras que los disponibles para cada locutor por separado son escasos, las mejores prestaciones deberían obtenerse aplicando adaptación a ambas variables, con MAP para el caso de tener muchos datos de adaptación y MLLR en el caso de datos muy limitados.

5.4 Conclusiones

La principal conclusión obtenida de los experimentos realizados es que diferenciar la generación de los modelos del reconocedor en función del locutor aporta beneficios en el proceso de reconocimiento, con lo que queda comprobado que las características de la señal de voz tienen dependencia con el locutor.

En cuanto a la adaptación MAP, indicar que se está consiguiendo un valor de 74.86 ± 0.32% para la tasa de aciertos, utilizando un valor de τ entre 10.1 y 10.5, adaptando sólo las medias, lo que representa una mejora relativa de la tasa de error de 6.47% respecto al experimento base (Resultado1).

Comparando estos resultados con la mejor configuración de adaptación MAP al género, es decir, utilizando el género del clasificador tanto en adaptación como en reconocimiento, cuya tasa de acierto es de 74.14 ± 0.32%, se está pasando de una mejora relativa de la tasa de error de 6.47% a 3.79%, lo que indica que introducir información del locutor en el reconocimiento de habla es mucho más beneficioso que introducir la información sobre el género.

Aunque la información sobre el locutor produce un incremento mayor en la tasa de aciertos que la información del género, las dos generan buenos resultados, por lo que se plantea la combinación de ambas. Utilizando MAP tanto en la adaptación al género como en la adaptación al locutor se está consiguiendo una tasa de acierto de 75.74 ± 0,32%, lo que representa una mejora relativa de la tasa de error de 9.75% respecto al experimento base (Resultado1), que verifica el hecho de que utilizar adaptación MAP de las dos informaciones conjuntamente proporciona mejores resultados que el uso de cualquiera de ellas por separado.

En cuanto a la técnica MLLR para realizar la adaptación al locutor, se ha comprobado que supera las prestaciones obtenidas con MAP. Esto era de esperar, ya que MLLR ofrece mejores

resultados que MAP cuando la cantidad de datos de adaptación es limitada, realizando una adaptación de todos los modelos, estén o no presentes en los datos.

Para el mejor caso de adaptación MLLR, utilizando las transcripciones reales de los datos de adaptación, se ha conseguido una mejora relativa de la tasa de error de 11.27% (tasa de aciertos de 76.15 ± 0.31%), incremento muy superior al obtenido para adaptación MAP al locutor (6.47%) e incluso superior al obtenido con adaptación MAP al género y al locutor (9.75%), concluyendo que las prestaciones de MLLR cuando los datos de adaptación son escasos, que es el caso del experimento actual, son muy buenas.

Comparando las dos técnicas MLLR utilizadas, una adaptación supervisada y una adaptación incremental no supervisada, ambas ofrecen resultados parecidos, 76.15 ± 0.31% para la primera y 76,08 ± 0.31% para la segunda, aunque hay que comentar que el tiempo necesario para realizar el reconocimiento de los locuciones de adaptación en la técnica no supervisada es muy elevado, debido sobre todo al aumento en el número de palabras del modelo de lenguaje.

Por último se ha evaluado la posibilidad de introducir adaptación MAP al género y MLLR al locutor. Utilizando la mejor configuración de la adaptación MAP al género y la forma supervisada de MLLR, se obtiene una tasa de aciertos de 76.73 ±0.31%, lo que supone una mejora relativa de la tasa de error de 13.43%, que supera en 2.16% al mayor incremento conseguido hasta el momento, que se correspondía con la adaptación MLLR al locutor (mejora relativa de la tasa de error de 11.27%).

Capítulo 6

Conclusiones y líneas de trabajo futuras

6.1 Conclusiones

El presente proyecto fin de carrera tiene como objetivo mejorar las prestaciones de un reconocedor automático de habla continua en castellano, adaptando sus características al género y al locutor.

Este reconocedor de habla fue implementado a partir de uno en inglés, desarrollado previamente en el Departamento de Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid.

Antes de comenzar el desarrollo, se realizó una batería de pruebas sobre el reconocedor en inglés, con el objetivo de conocer cómo modificar distintos parámetros de configuración y obtener

información de la influencia de los mismos sobre los resultados del reconocimiento. Entre las conclusiones obtenidas se tiene que dichos resultados mejoran:

- Con el aumento de la frecuencia de muestreo de la base de datos.
- Al aplicar la técnica CMN en la parametrización, reduciéndose la influencia de los micrófonos.
- Al utilizar los coeficientes de aceleración en la parametrización, aunque como contrapartida provoca un aumento significativo de la información a tratar.
- Al utilizar HMMs diferentes para modelar las transiciones entre palabras, ya que estas transiciones pueden ser de naturaleza muy distinta.
- Con la utilización de trifonemas en lugar de monofonos. Los resultados mejoran cuanto más elaborados son los modelos de trifonemas, aunque dicha mejora empieza a reducirse al utilizar mezclas con un número elevado de gausianas, lo que podría estar produciendo una sobreadaptación a los datos de entrenamiento.

Por último se observó el efecto de la penalización por inserción de palabra y del peso del modelo de lenguaje. A medida que el primero toma valores más negativos, lo que penaliza más la inserción de nuevas palabras, el número de inserciones disminuye, aumentando también el número de eliminaciones. En cuanto al peso del modelo de lenguaje, a medida que aumenta se le da mayor importancia al modelo de lenguaje. Si es demasiado elevado puede llegar a anular la importancia de la probabilidad acústica del modelo, lo que se ve reflejado en un aumento de las eliminaciones. Debido a que el número de inserciones y de eliminaciones varían de forma inversa al modificar estos parámetros, lo óptimo resulta elegir un valor para ellos donde el número de inserciones y de eliminaciones se igualen, lo que aporta los mejores valores para la tasa de acierto.

Una vez adquirido el conocimiento necesario a partir de las pruebas anteriores, se procedió al desarrollo del reconocedor automático de habla en castellano utilizado en el presente proyecto fin de carrera. Para ello, se utilizó la base de datos de habla leída MICROAES. Con este sistema base se obtuvo una tasa de acierto de 73.12%, que se pretendió mejorar introduciendo información del género y del locutor.

Se plantearon distintas pruebas para comprobar la influencia sobre los resultados de reconocimiento de la utilización del género de las locuciones. En primer lugar se utilizó la adaptación MAP al género. Tras obtener un conjunto de modelos bien entrenados (que corresponden con los obtenidos en la implementación de referencia), se adaptaron dichos modelos utilizando los datos de adaptación de cada género (que se corresponden con los datos de entrenamiento separados en función del género de cada locución). De esta forma se obtuvieron dos nuevos conjuntos de modelos, uno representativo de los locutores masculinos y otro de los femeninos.

Tras distintas pruebas, realizando adaptación MAP de medias o de medias y varianzas, utilizando el género real de las locuciones o el clasificador de género para decidirlo, y modificando el parámetro que da mayor o menor importancia a la probabilidad de los datos a priori en dicha adaptación, se consiguió una mejora relativa de la tasa de error de 3.79%, respecto al experimento de referencia, obteniéndose estos resultados con un valor de τ igual a 1, adaptando sólo las medias, y utilizando el género decidido por el clasificador tanto para obtener los datos de adaptación como para elegir el conjunto de modelos a usar en el reconocimiento.

Como a priori la cantidad de datos de adaptación de cada género es abundante, se planteó el entrenamiento completo de los modelos separados por género, obteniéndose los mejores resultados utilizando el clasificador de género tanto al separar los datos de entrenamiento como al decidir el conjunto de modelos a utilizar en el reconocimiento. La mejora de la tasa relativa de error respecto al experimento base fue del 6.58%.

Las conclusiones obtenidas de estos experimentos son:

- Queda probado que las características de la señal de voz tienen dependencia con el género, ya que diferenciar la generación de los modelos del reconocedor en función del género de los locutores aporta beneficios en el proceso de reconocimiento.
- Se produce una mejora significativa de los resultados al introducir la adaptación MAP al género, pero se obtienen mejores resultados entrenando los modelos desde el principio, lo que demuestra que los datos de entrenamiento son suficientemente abundantes al separarlos en locuciones masculinas y femeninas.
- Utilizar el clasificador de género permite aprovechar positivamente el hecho de que locuciones de un género tengan más características comunes con el género contrario. Utilizándolo en entrenamiento, se consigue modelar cada género con muestras de similares características, aunque se correspondan con géneros diferentes en la realidad. En la fase de test del reconocedor, permite utilizar los modelos que mejor se corresponden con las características acústicas de las muestras de voz de la locución, lo que producirá transcripciones más exactas.

En cuanto a las pruebas planteadas para comprobar la influencia sobre los resultados de reconocimiento de la información del locutor, se propuso utilizar una adaptación MAP y una adaptación MLLR, esta última porque a priori ofrece mejores resultados que la primera en el caso de tener pocos datos de adaptación.

Tras obtener un conjunto de modelos bien entrenados, que coincide con los obtenidos en la implementación de referencia, se adaptaron dichos modelos utilizando los datos correspondientes a cada locutor. Con dicha adaptación se obtuvieron:

- 300 conjuntos de modelos nuevos, cada uno representativo de cada locutor, para el caso de adaptación MAP.
- 300 matrices de transformación del conjunto inicial, cada una representativa de cada locutor, para el caso de MLLR. Mencionar que sólo se obtiene una matriz por locutor porque se ha configurado que todos los modelos sufran la misma transformación.

En cuanto a la adaptación MAP al locutor, se consiguió una mejora relativa de la tasa de error respecto al experimento de referencia de 6.47%, lo que supera en un 2.68% la mejora máxima producida para adaptación MAP al género.

En cuanto a la adaptación MLLR, se evaluó una adaptación supervisada y una adaptación incremental no supervisada, obteniéndose los mejores resultados en el primer caso, posiblemente por el error de reconocimiento introducido en el caso no supervisado al obtener las transcripciones de adaptación. Con esta técnica se consiguió una mejora relativa de la tasa de error respecto al experimento base de 11.27%, un 4.8% más que para adaptación MAP al locutor.

La principal conclusión obtenida de los experimentos realizados es que diferenciar la generación de los modelos del reconocedor en función del locutor aporta beneficios en el proceso de reconocimiento, con lo que queda comprobado que las características de la señal de voz tienen dependencia con el locutor. Además, debido a que MLLR aporta mayores beneficios que MAP, queda comprobado que los datos de adaptación son escasos, ya que la técnica MLLR ofrece mejores prestaciones que MAP cuando se produce este hecho. Por esta razón, el entrenamiento desde el principio de los modelos separados por locutor no aportó buenos resultados.

Aunque la información sobre el locutor produce un incremento mayor en la tasa de aciertos que la información del género, las dos generan buenos resultados, por lo que se planteó la combinación de ambas.

Utilizando MAP tanto en la adaptación al género como en la adaptación al locutor se consiguió una mejora relativa de 9.75%, valor muy superior al obtenido con adaptación MAP al género (3.79%) o al locutor (6.47%), lo que verifica que utilizar adaptación MAP de las dos informaciones conjuntamente proporciona mejores resultados que el uso de cualquiera de ellas por separado.

Sin embargo, se obtienen peores resultados que con la adaptación MLLR al locutor, donde se conseguía una mejora relativa de la tasa de error de 11.27%. Utilizando adaptación al género con MAP y adaptación al locutor con MLLR se consigue la mejor combinación, proporcionando

una mejora relativa de la tasa de error de 13.43%, que supera en 2.16% al mayor incremento conseguido hasta el momento, que se correspondía con la adaptación MLLR al locutor.

6.2 Líneas de trabajo futuras

Este proyecto fin de carrera surge de la idea de realizar un reconocedor automático de habla espontánea en castellano. Como ya se ha comentado, debido a la escasez de datos de entrenamiento de la base de datos de habla espontánea, se realizó un reconocedor de habla continua, y se adaptó éste a la información de dicha base de datos. Esta adaptación se llevó a cabo en un proyecto paralelo [Alc07]. Mientras, en este proyecto se investigaron técnicas para introducir información sobre el género o el locutor en el reconocedor de habla continua, obteniéndose aquellas que producen una mejora significativa de los resultados. Como línea principal de trabajo futuro surge la necesidad de combinar las técnicas aplicadas para adaptar el reconocedor al habla espontánea junto con aquellas de adaptación al género y al locutor que mejores beneficios han aportado sobre el reconocedor de habla continua, comprobando así si esta información influye de igual manera sobre el reconocedor de habla espontánea, y como varían sus resultados iniciales.

En cuanto a aspectos relacionados con la implementación realizada sobre la adaptación MLLR, en este proyecto se ha utilizado una única transformación común para todas las medias de todas las gausianas de todos los HMMs del sistema. MLLR permite introducir unos árboles de regresión, de forma que se puedan realizar grupos de modelos (que compartan alguna característica) obteniéndose transformaciones diferentes para cada uno de ellos. Se puede plantear la utilización de estas agrupaciones, puesto que si se dispone de la cantidad de datos suficientes, deberían producir mejoras respecto al uso de un único grupo. Para ello será necesario realizar un estudio sobre las características a utilizar para realizar dichas agrupaciones, y sobre los criterios que indiquen el número óptimo de grupos en función de los datos disponibles.

Si se tienen en cuenta la base de datos utilizada para el reconocedor de habla continua (MICROAES), ésta ofrece grabaciones para 4 micrófonos diferentes: 2 situados cerca del locutor, y otros dos situados a una distancia superior. Para los estudios llevados a cabo, sólo se ha utilizado la información de los dos micrófonos más cercanos, de forma que la voz utilizada fuera de la mayor calidad posible, con una influencia mínima del ruido externo. Ya que se dispone de datos de menor calidad, con mayor influencia de factores externos debido a que los micrófonos tenían una situación más alejada del locutor, sería interesante observar el efecto producido al

introducir esta información en el sistema, comprobando si influye de la misma manera que el resto de información, o si introduce algún tipo de ventaja o inconveniente.

Relacionado también con la base de datos MICROAES, mencionar que el modelo de lenguaje utilizado es una combinación equiprobable de las palabras de test de dicha base de datos. Además, para el caso de adaptación MLLR no supervisada fue necesario utilizar un nuevo modelo de lenguaje, combinando todas las palabras de los datos de adaptación. Sería muy beneficioso trabajar con un modelo de lenguaje genérico, que haga uso de reglas que impidan el reconocimiento de frases sin sentido. Además, esto permitiría una aplicación real del reconocedor, donde las palabras a reconocer no pertenezcan al grupo limitado de palabras de test del sistema actual.

Por último, mencionar que aunque los HMMs están muy extendidos como técnica de reconocimiento de patrones, se puede plantear el uso de otras técnicas, como la combinación de HMMs con redes neuronales, evaluando tanto la complejidad técnica como computacional de esta combinación, y si las mejoras obtenidas compensan la complejidad introducida en el sistema.

Capítulo 7

Presupuesto

7.1 Introducción

En este capítulo se va a calcular de manera aproximada el coste de la realización del proyecto. Los costes directos del mismo se van a detallar divididos en dos grupos: coste del equipo utilizado y coste del personal que ha intervenido.

Para todos los cálculos realizados se considerará que el tiempo necesario para la realización del proyecto ha sido de 1 año.

7.2 Coste del material

Los recursos materiales que se han utilizado para la realización del proyecto se pueden dividir en:

- Ordenador personal.
- Recursos computacionales del departamento.
- Base de datos.
- Puesto de trabajo.

En cuanto al puesto de trabajo, se va a suponer que sólo se necesita un puesto físico para el ingeniero que desarrolla el proyecto, mientras que el director del mismo no tiene necesidad de lugar físico de trabajo.

Si para cubrir la necesidad anterior se alquila una oficina para 4 puestos de trabajo, suponiendo que los otros 3 puestos se utilizan en diferentes proyectos, sólo 1 de ellos es computable para el presupuesto actual. Si el coste mensual de la oficina es de 1000€ (IVA incluido), el coste anual por puesto asciende a 3000€ (IVA incluido).

En cuanto al software necesario, se tiene:

- Conjunto de herramientas de HTK: sin coste asociado.
- Base de datos MICROAES: con un precio de 18000€.

Al coste de la base de datos se le va a aplicar una amortización, puesto que este recurso podrá ser utilizado en otros proyectos. Suponiendo 60 meses como periodo de depreciación, como se ha utilizado durante 12 meses, a este proyecto se le imputará un coste de 3600€ por el uso de la base de datos MICROAES.

En cuanto al hardware, se utiliza un ordenador personal y unos recursos computacionales del Departamento de Teoría de la Señal y Comunicaciones.

El ordenador personal tampoco será de uso exclusivo para el proyecto, por lo que suponiendo también un periodo de depreciación de 60 meses para un ordenador cuyo precio aproximado es de 1500€ (IVA incluido), el coste asociado a este recurso por los 12 meses en los que se ha utilizado es de 300€.

Para el cálculo del coste de los recursos computacionales del Departamento de Teoría de la Señal y Comunicaciones, hay que tener en cuenta que sólo se usaron para pruebas puntuales durante 2 meses.

En el coste total de este sistema debe considerarse tanto el precio del hardware como el mantenimiento que requiere. Suponiendo que estos recursos tienen un precio de 150000€, con un coste de mantenimiento de 22000€ anules, si como periodo de depreciación se consideran 60 meses, se necesita un mantenimiento durante esos 5 años, con lo que el precio total asciende a 260000€.

Teniendo en cuenta la amortización, 2 meses de uso exclusivo del recurso anterior tiene un coste de 8666.67€. Como el sistema soporta simultáneamente 30 usuarios, el coste asociado a su uso por parte del ingeniero de este proyecto es de 288.89€.

La Tabla 8 engloba todos los costes de material considerados:

Recurso	Precio (€)
Puesto de trabajo	3000
Base de datos MICROAES	3600
Ordenador	300
Recurso computacional compartido	288.89

Tabla 8.- Costes de material

7.3 Coste del personal

El personal necesario para la realización del proyecto ha sido un Ingeniero de Telecomunicaciones y un Director del proyecto.

Para el cálculo de los costes del Ingeniero que se encarga del desarrollo, se tiene en cuenta que va a tener dedicación plena durante 12 meses, y que el precio de 1 mes de trabajo de un Ingeniero es de 2694.39 € (IVA incluido). Por lo tanto, el coste asociado es de 32332.68 €.

En cuanto al director, se supone una dedicación del 10 % del tiempo total del proyecto y que el precio de 1 mes de trabajo de un Ingeniero Senior es de 4289.54 € (IVA incluido). Por lo tanto, el coste asociado es de 5147.45 €.

La Tabla 9 engloba todos los costes de personal considerados:

Recurso	Precio (€)
Ingeniero	32332.68
Ingeniero Senior	5147.45

Tabla 9.- Costes de personal

7.4 Presupuesto total

Para el cálculo del presupuesto total, además de tener en cuenta la suma de los costes asociados al material y al personal, se introducirá un nuevo factor asociado a costes indirectos que se puedan producir, por un valor del 20% de la suma anterior.

En la Tabla 8 se desglosa el coste asociado al material, que en total asciende a 7188.89 €.

De igual forma, en la Tabla 9 se desglosa el coste asociado al personal, que en total asciende a 37480.13 €.

Los costes indirectos se calculan como un 20% de la suma anterior, ascendiendo a 8933.80€.

Teniendo en cuenta los tres factores anteriores, el presupuesto total de este proyecto asciende a la cantidad de 53602.82 € (IVA incluido).

Referencias

[YEG+06] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P.: 'The HTK Book (for HTK Version 3.4)'. <http://htk.eng.cam.ac.uk/>.

[HAH01] Huang, X., Acero, A., and Hon, H-W.: 'Spoken Language Processing. A guide toTheory, Algorithm, and System Development'. Prentice Hall PTR, 2001.

[GL94] Gauvain, J-L., and Lee, C-H.: 'Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains', IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 2, April 1994.

[Mar06] Martín Iglesias, D., Informe interno sobre 'Verificador de Locutor' desarrollado en el Grupo de Procesado Multimedia del Departamento de Teoría de la Señal y Comunicaciones, Universidad Carlos III de Madrid. 2006.

[RQD00] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B.: 'Speaker Verification Using Adapted Gaussian Mixture Models', Digital Signal Processing 10, 9-41, 2000.

[RR95] Reynolds, D. A., and Rose, R. C.: 'Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models', IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, January 1995.

[Gal98] Gales, M. J. F.: 'Maximum likelihood linear transformations for HMM-based speech recognition', Computer Speech and Language 12, 75-98, Cambridge University Engineering Department, 1998.

[Alc07] Alcón Paniagua, S.: 'Diseño de un reconocedor automático de habla espontánea en castellano'. Proyecto fin de carrera, Universidad Carlos III de Madrid, 2007.

[Fer03] Ferreiros, J.: '¿Qué queremos que sea Tecnología del Habla?', XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. Procesamiento del Lenguaje Natural 31: 375-380, Universidad de Alcalá, 10, 11 y 12 de septiembre de 2003. <http://www.sepln.org/revistaSEPLN/revista/31/31-Pag375.pdf>

[Her03] Hernández, L.: 'Modelo de evolución de la tecnología del habla, y tendencias futuras'. XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. Procesamiento del Lenguaje Natural 31: 369-374, Universidad de Alcalá, 10, 11 y 12 de septiembre de 2003. <http://www.sepln.org/revistaSEPLN/revista/31/31-Pag369.pdf>

[CEIDIS]http://www.ceidis.ula.ve/cursos/humanidades/fonetica/tutorial_de_linguistica/techabla.html, accedida en septiembre de 2010.

[Lli09] Llisterri, J.:
<http://liceu.uab.es/~joaquim/language_technology/HLT/tecnol_ling_habla.html>

[ML96] Moure, T. and Llisterri, J.: 'Lenguaje y nuevas tecnologías. El campo de la lingüística computacional', en Fernández Pérez, M (Coord.) *Avances en lingüística aplicada*. Santiago de Compostela: Universidad de Santiago de Compostela, Servicio de Publicación e Intercambio Científico (Avances en, 4). pp. 147-228. 1996
<http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html>

[REV] <http://reeducacionvocal.blogspot.com>

[Miy04] Miyara, F.: 'La voz humana'. Monografía para la Cátedra de Procesamiento Digital de Señales de Voz, Universidad Nacional de Rosario, Argentina, 2004.

[FUR89] Furui, S.: 'Digital Speech Processing, Synthesis, and Recognition'. Marcel Dekker. 1989.

[RJ93] Rabiner, L. R., and Juang, B.H.: 'Fundamental of Speech Recognition'. Prentice-Hall, 1993.

[EUMUS] <http://www.eumus.edu.uy/docentes/maggiolo/acuapu/sap.html>

[Col01] Colás Pasamontes, J.: 'Estrategias de Incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español', Capítulo 2, 4, 2001. <http://elies.rediris.es/elies12/>

[WSJ0] Base de datos WSJ0 (Wall Street Journal), ARPA Spoken Language Program, 1991. <http://catalog.elra.info/product_info.php?products_id=695, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S6A>

[Vic05] de Vicente Peña, J.: 'Speech recognition with the Wall Street Journal database'. KTH University, Stockholm, and Universidad Carlos III de Madrid. 2005.

[CMUv06] versión 0.6 del diccionario CMU (Carnegie Mellon University Pronouncing Dictionary). <http://www.speech.cs.cmu.edu/cgi-bin/pronounce>

[MIC04] Base de datos MICROAES, ATLAS (Applied Technologies on Language and Speech S.L.), 2004. <http://www.elda.org/catalogue/en/speech/S0165.html, http://catalog.elra.info/search_result.php?keywords=microaes&language=en>

[COE05] Santiago Rodríguez y Jesús Carretero, grupo de trabajo COES, Departamento de Arquitectura y Tecnología de Sistemas Informáticos (DATSI), Universidad Politécnica de Madrid (UPM), 2005. <http://www.datsi.fi.upm.es/~coes/espell_leame/espell_leame.html>.

Anexo I

En este anexo se van a incluir las tablas correspondientes a algunos experimentos realizados.

Tabla 10.- Clasificador de género, grupos independientes del locutor, 1 reestimación

Número de gausianas	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%).
2	64.00	99.65	81.82
4	85.41	98.71	92.06
8	88.82	96.59	92.71
16	90.24	95.18	92.71
32	92.12	93.18	92.65
64	93.29	92.59	92.94
128	94.24	92.12	93.18
256	95.65	92.00	93.82
512	96.00	92.00	94.00

Tabla 11.- Clasificador de género, grupos independientes del locutor, 2 reestimaciones

Número de gausianas	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%)
2	84.94	97.65	91.29
4	92.00	94.59	93.29
8	90.82	93.88	92.35
16	92.82	93.06	92.94
32	94.12	92.47	93.29
64	95.18	92.35	93.76
128	96.00	92.00	94.00
256	96.00	92.00	94.00
512	96.00	92.00	94.00

Tabla 12.- Clasificador de género, grupos independientes del locutor, 3 reestimaciones

Número de gausianas	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%)
2	86.94	95.29	91.12
4	90.12	94.35	92.24
8	90.94	93.53	92.24
16	92.59	92.82	92.71
32	94.00	92.24	93.12
64	95.88	92.00	93.94
128	96.00	92.00	94.00
256	96.00	92.00	94.00
512	96.00	92.00	94.00

Tabla 13.- Clasificador de género, grupos independientes del locutor, ajuste del umbral

Umbral	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%)
0	95.88	92.00	93.94
10	95.76	92.00	93.88
100	95.76	92.00	93.88
200	95.18	92.24	93.71
300	94.47	92.59	93.53
400	93.88	92.94	93.41
500	93.18	93.41	93.29
600	92.71	94.71	93.71
700	92.24	95.29	93.76
800	91.18	96.24	93.71
900	89.76	96.82	93.29
450	93.41	93.18	93.29
550	92.94	93.76	93.35
460	93.29	93.18	93.24
470	93.29	93.18	93.24
480	93.29	93.29	93.29
490	93.18	93.41	93.29
1000	88.24	98.24	93.24

Tabla 14.-Clasificador de género, grupos dependientes del locutor,3 reestimaciones

Número de gausianas	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%)
2	87.07	95.26	91.25
4	90.14	95.92	93.08
8	91.84	96.24	94.08
16	93.20	96.57	94.92
32	94.56	97.06	95.83
64	94.90	97.71	96.33
128	95.24	97.88	96.58
256	95.24	98.04	96.67
512	95.41	98.37	96.92

Tabla 15.- Clasificador de género, grupos dependientes del locutor, ajuste del umbral

Umbral	Tasa de acierto masculinos (%)	Tasa de acierto femeninos (%)	Tasa de acierto común (%)
450	94.24	93.76	94.00
460	94.00	93.88	93.94
470	94.00	94.00	94.00
480	93.88	94.12	94.00
490	93.88	94.12	94.00
500	93.88	94.24	94.06

Tabla 16.- Barrido de τ, utilizando género real en la fase de entrenamiento y género del clasificador en reconocimiento. Se adaptan medias y varianzas.

T	Tasa de acierto (%)
0.1	73.43
1	73.58
5	73.81
10	73.97
20	73.97
40	73.93
60	73.88
80	73.83
100	73.77
200	73.59
2000	73.22

Tabla 17.- Barrido de τ, utilizando género del clasificador en la fase de entrenamiento y género del clasificador en reconocimiento. Sólo se adaptan las medias.

T	Tasa de acierto (%)
0.1	74.11
1	74.14
5	74.13
10	74.10
20	73.97
40	73.87
60	73.82
80	73.75
100	73.68
200	73.51
2000	73.51

Tabla 18.- Barrido de τ, realizando adaptación MAP al locutor. Sólo se adaptan las medias.

T	Tasa de acierto (%)
10	74,55
10,01	74,86
10,02	74,86
10,025	74,86
10,03	74,86
10,04	74,86
10,05	74,86
10,075	74,85
10,1	74,85
10,15	74,84
10,2	74,83
10,3	74,83
10,4	74,82
10,5	74,82

10,6	74,82
10,7	74,81
10,8	74,80
10,9	74,79
11	74,80
11,1	74,78
11,2	74,79
11,3	74,79
11,4	74,79
11,5	74,78
11,6	74,78
11,7	74,78
11,8	74,78
11,9	74,79
12	74,79
13	74,76
14	74,74
15	74,67
16	74,64
17	74,62
18	74,61
19	74,59
20	74,55
30	74,34
40	74,28
50	74,17
60	74,07
70	74,01
80	73,93
90	73,88

100	73,83
200	73,61
300	73,49
400	73,42
500	73,37
600	73,33

Tabla 19.- Barrido de τ, realizando adaptación MAP al locutor y adaptación MAP al género. Sólo se adaptan las medias

T	Tasa de acierto (%)
1	75,47
5	75,74
10	75,51
10.01	75,51
10.05	75,51
10.1	75,51
10.5	75,48
11	75,47